Formemes in English-Czech Deep Syntactic MT *

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel,
Martin Majliš, Michal Novák, and **David Mareček**
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague
{odusek,zabokrtsky,popel,majlis,mnovak,marecek}@ufal.mff.cuni.cz

Abstract

One of the most notable recent improvements of the TectoMT English-to-Czech translation is a systematic and theoretically supported revision of *formemes*—the annotation of morpho-syntactic features of content words in deep dependency syntactic structures based on the Prague tectogrammatics theory. Our modifications aim at reducing data sparsity, increasing consistency across languages and widening the usage area of this markup. Formemes can be used not only in MT, but in various other NLP tasks.

1 Introduction

The cornerstone of the TectoMT tree-to-tree machine translation system is the deep-syntactic language representation following the Prague tectogrammatics theory (Sgall et al., 1986), and its application in the Prague Dependency Treebank (PDT) 2.0[1] (Hajič et al., 2006), where each sentence is analyzed to a dependency tree whose nodes correspond to content words. Each node has a number of attributes, but the most important (and difficult) for the transfer phase are *lemma*—lexical information, and *formeme*—surface morpho-syntactic information, including selected auxiliary words (Ptáček and Žabokrtský, 2006; Žabokrtský et al., 2008).

This paper focuses on formemes—their definition and recent improvements of the annotation, which has been thoroughly revised in the course of preparation of the CzEng 1.0 parallel corpus (Bojar et al., 2012b), whose utilization in TectoMT along with the new formemes version has brought the greatest benefit to our English-Czech MT system in the recent year. However, the area of possible application of formemes is not limited to MT only or to the language pair used in our system; the underlying ideas are language-independent.

We summarize the development of morpho-syntactic annotations related to formemes (Section 2), provide an overview of the whole TectoMT system (Section 3), then describe the formeme annotation (Section 4) and our recent improvements (Section 5), as well as experimental applications, including English-Czech MT (Section 6). The main asset of the formeme revision is a first systematic reorganization of the existing practical aid, providing it with a solid theoretical base, but still bearing its intended applications in mind.

2 Related Work

Numerous theoretical approaches had been made to morpho-syntactic description, mainly within valency lexicons, starting probably with the work by Helbig and Schenkel (1969). Perhaps the best one for Czech is PDT-VALLEX (Hajič et al., 2003), listing all possible subtrees corresponding to valency arguments (Urešová, 2009). Žabokrtský (2005) gives an overview of works in this field.

* This research has been supported by the grants FP7-ICT-2009-4-247762 (FAUST), FP7-ICT-2009-4-249119 (Metanet), LH12093 (Kontakt II), DF12P01OVV022 (NAKI), 201/09/H057 (Czech Science Foundation), GAUK 116310, and SVV 265 314. 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).

[1] http://ufal.mff.cuni.cz/pdt2.0

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 267–274,
Montréal, Canada, June 7-8, 2012.

This kind of information has been most exploited in structural MT systems, employing semantic relations (Menezes and Richardson, 2001) or surface tree substructures (Quirk et al., 2005; Marcu et al., 2006). Formemes, originally developed for Natural Language Generation (NLG) (Ptáček and Žabokrtský, 2006), have been successfully applied to MT within the TectoMT system. Our revision of formeme annotation aims to improve the MT performance, keeping other possible applications in mind.

3 The TectoMT English-Czech Machine Translation System

The TectoMT system is a structural machine translation system with deep transfer, first introduced by Žabokrtský et al. (2008). It currently supports English-to-Czech translation. Its analysis stage follows the Prague tectogrammatics theory (Sgall, 1967; Sgall et al., 1986), proceeding over two layers of structural description, from shallow (*analytical*) to deep (*tectogrammatical*) (see Section 3.1).

The transfer phase of the system is based on Maximum Entropy context-sensitive translation models (Mareček et al., 2010) and Hidden Tree Markov Models (Žabokrtský and Popel, 2009). It is factorized into three subtasks: lemma, formeme and grammatemes translation (see Sections 3.2 and 3.3).

The subsequent generation phase consists of rule-based components that gradually change the deep target language representation into a shallow one, which is then converted to text (cf. Section 6.1).

The version of TectoMT submitted to WMT12[2] builds upon the WMT11 version. Several rule-based components were slightly refined. However, most of the effort was devoted to creating a better and bigger parallel treebank—CzEng 1.0[3] (Bojar et al., 2012b), and re-training the statistical components on this resource. Apart from bigger size and improved filtering, one of the main differences between CzEng 0.9 (Bojar and Žabokrtský, 2009) (used in WMT11) and CzEng 1.0 (used in WMT12) is the revised annotation of formemes.

[2] http://www.statmt.org/wmt12

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

3.1 Layers of structural analysis

There are two distinct structural layers used in the TectoMT system:

- *Analytical layer*. A surface syntax layer, which includes all tokens of the sentence, organized into a labeled dependency tree. The labels correspond to surface syntax functions.

- *Tectogrammatical layer*. A deep syntax/semantic layer describing the linguistic meaning of the sentence. Its dependency trees include only content words as nodes, assigning to each of them a deep lemma (*t-lemma*), a semantic role label (*functor*), and other deep linguistic features (*grammatemes*), such as semantic part-of-speech, person, tense or modality.

The analytical layer can be obtained using different dependency parsers (Popel et al., 2011); the tectogrammatical representation is then created by rule-based modules from the analytical trees.

In contrast to the original PDT annotation, the TectoMT tectogrammatical layer also includes *formemes* describing the surface morpho-syntactic realization of the nodes (cf. also Section 3.3).

3.2 Transfer: Translation Factorization and Symmetry

Using the tectogrammatical representation in structural MT allows separating the problem of translating a sentence into relatively independent simpler subtasks: lemma, functors, and grammatemes translation (Bojar et al., 2009; Žabokrtský, 2010). Since topology changes to deep syntax trees are rare in MT transfer, each of these three subtasks allows a virtually symmetric source-target one-to-one mapping, thus simplifying the initial n-to-m mapping of word phrases or surface subtrees.

Žabokrtský et al. (2008) obviated the need for transfer via functors (i.e. semantic role detection) by applying a formeme transfer instead. While formeme values are much simpler to obtain by automatic processing, this approach preserved the advantage of symmetric one-to-one value translation.

Moreover, translations of a given source morpho-syntactic construction usually follow a limited number of patterns in the target language regardless of

their semantic functions, e.g. a finite clause will most often be translated as a finite clause.

3.3 Motivation for the Introduction of Formemes

Surface-oriented formemes have been introduced into the semantics-oriented tectogrammatical layer, as it proves beneficial to combine the deep syntax trees, smaller in size and more consistent across languages, with the surface morphology and syntax to provide for a straightforward transition to the surface level (Žabokrtský, 2010).

The three-fold factorization of the transfer phase (see Section 3.2) helps address the data sparsity issue faced by today's MT systems. As the translation of lemmas and their morpho-syntactic forms is separated, combinations unseen in the training data may appear on the output.

To further reduce data sparsity, only minimal information needed to reconstruct the surface form is stored in formemes; morphological categories derivable from elsewhere, i.e. morphological agreement or grammatemes, are discarded.

4 Czech and English Formemes in TectoMT

A *formeme* is a concise description of relevant morpho-syntactic features of a node in a tectogrammatical tree (deep syntactic tree whose nodes usually correspond to content words). The general shape of revised Czech and English formemes, as implemented within the Treex[4] NLP framework (Popel and Žabokrtský, 2010) for the TectoMT system, consists of three main parts:

1. *Syntactic part-of-speech.*[5] The number of syntactic parts-of-speech is very low, as only content words are used on the deep layer and the categories of pronouns and numerals have been divided under nouns and adjectives according to syntactic behavior (Ševčíková-Razímová and Žabokrtský, 2006). The possible values are `v` for verbs, `n` for nouns, `adj` for adjectives, and `adv` for adverbs.
2. *Subordinate conjunction/preposition.* Applies only to formemes of prepositional phrases and subordinate clauses introduced by a conjunction and contains the respective conjunction or preposition; e.g. `if`, `on` or `in_case_of`.
3. *Form.* This part represents the morpho-syntactic form of the node in question and depends on the part-of-speech (see Table 1).

The two or three parts are concatenated into a human-readable string to facilitate usage in hand-written rules as well as statistical systems (Žabokrtský, 2010), producing values such as `v:inf`, `v:if+fin` or `n:into+X`. Formeme values of nodes corresponding to uninflected words are atomic.

Formemes are detected by rule-based modules operating on deep and surface trees. Example deep syntax trees annotated with formemes are shown in Fig. 1. A listing of all possible formeme values is given in Table 1.

Verbal formemes remain quite consistent in both languages, except for the greater range of forms in English (Czech uses adjectives or nouns instead of gerunds and verbal attributes). Nominal formemes differ more significantly: Czech is a free-word order language with rich morphology, where declension is important to syntactic relations—case is therefore included in formemes. As English makes its syntactic relations visible rather with word-order than with morphology, English formemes indicate the syntactic position instead. The same holds for adjectival complements to verbs. Posession is expressed mostly using nouns in English and adjectives in Czech, which is also reflected in formemes.

5 Recent Markup Improvements

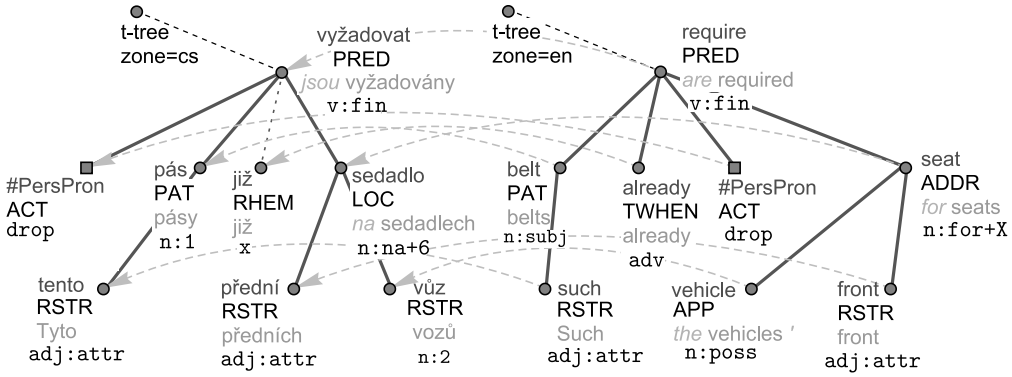
Our following markup innovations address several issues found in the previous version and aim to adapt the range of values more accurately to the intended applications.

5.1 General Form Changes

The relevant preposition and subordinate conjunction nodes had been selected based on their dependency labels; we use a simple part-of-speech tag filter instead in order to minimize the influence of parsing errors and capture more complex prepositions,

[4] `http://ufal.mff.cuni.cz/treex/`, `https://metacpan.org/module/Treex`

[5] Cf. Section 5.2 for details.



[en] Such belts already are required for the vehicles' front seats. [cs] Tyto pásy jsou již vyžadovány na předních sedadlech vozů.

Figure 1: An example English and Czech deep sentence structure annotated with formemes (in typewriter font).

Formeme	Language	Definition
`v:(P+)fin`	both	Verbs as heads of finite clauses
`v:rc`	both	Verbs as heads of relative clauses
`v:(P+)inf`	both	Infinitive clauses; typically with the particle *to* in English*
`v:(P+)ger`	EN	Gerunds, e.g. *I like reading (`v:ger`), but I am tired of arguing (`v:of+ger`).*
`v:attr`	EN	Present or past participles (i.e. *-ing* or *-ed* forms) in the attributive syntactic position, e.g. *Striking (`v:attr`) teachers hate bored (`v:attr`) students.*
`n:[1..7]`	CS	Bare nouns; the numbers indicate morphological case†
`n:X`	CS	Bare nouns that cannot be inflected
`n:subj`	EN	Nouns in the subject position (i.e. in front of the main verb of the clause)
`n:obj`	EN	Nouns in the object position (i.e. following the verb with no preposition)
`n:obj1, n:obj2`	EN	Nouns in the object position; distinguishing the two objects of ditransitive verbs (e.g. *give*, *consider*)
`n:adv`	EN	Nouns in an adverbial position, e.g. *The sales went up by 1 % last month*
`n:P+X`	EN	Prepositional phrases
`n:P+[1..7]`	CS	Prepositional phrases; the preposition surface form is combined with the required case‡
`n:attr`	both	Nominal attributes, e.g. *insurance company* or *president Smith* in English and *prezident Smith* in Czech
`n:poss`	EN	English possessive pronouns and nouns with the *'s* suffix
`adj:attr`	both	Adjectival attributes (Czech inflection forms need not be stored thanks to congruency with the parent noun)
`adj:compl`	EN	Direct adjectival complements to verbs
`adj:[1..7]`	CS	Direct adjectival complements to verbs (morphological case must be stored in Czech, as it is determined by valency)
`adj:poss`	CS	Czech possesive adjectives and pronouns; a counterpart to English `n:poss`
`adv`	both	Adverbs (not inflected, can take no prepositions etc.)
`x`	both	Coordinating conjunctions, other uninflected words
`drop`	both	Deep tree nodes which do not appear on the surface (e.g. pro-drop pronouns)

*I.e. infinitives as head of clauses, not infinitives as parts of compound verb forms with finite auxiliary verbs.

†Numbers are traditionally used to mark morphological case in Czech; `1` stands for nominative, `2` for genitive etc.

‡Since many prepositions may govern multiple cases in Czech, the case number is necessary.

Table 1: A listing of all possible formeme values, indicating their usage in Czech, English or both languages. "*P+*" denotes the (lowercased) surface form of a preposition or a subordinate conjunction. Round brackets denote optional parts, square brackets denote a set of alternatives.

e.g. *in case of*. Our revision also allows combining prepositions with all English gerunds and infinitives, preventing a loss of important data.

We also use the lowercased surface form in the middle formeme part instead of lemmas to allow for a more straightforward surface form generation.

5.2 Introducing Syntactic Part-of-Speech

Formemes originally contained the semantic part-of-speech (sempos) (Razímová and Žabokrtský, 2006) as their first part. We replaced it with a *syntactic* part-of-speech (syntpos), since it proved complicated to assign sempos reliably by a rule-based module and morpho-syntactic behavior is more relevant to formemes than semantics.

The syntpos is assigned in two steps:

1. A preliminary syntpos is selected, using our categorization based on the part-of-speech tag and lemma.

2. The final syntpos is selected according to the syntactic position of the node, addressing nominal usage of adjectives and cardinal numerals (see Sections 5.4 and 5.5).

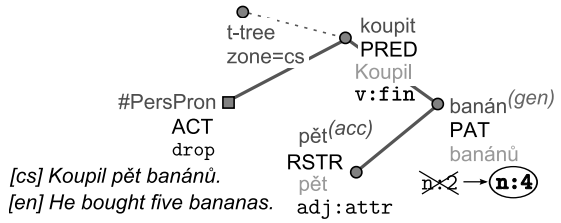
5.3 Capturing Czech Nominal Attributes

Detecting the attributive usage of nouns is straightforward for English, where any noun depending directly on another noun is considered an attribute. In Czech, one needs to distinguish case-congruent attributes from others that have a fixed case. We aimed at assigning the `n:attr` formeme only in the former case and thus replaced the original method based on word order with a less error-prone one based on congruency and named entity recognition.

5.4 Numerals: Distinguishing Usage and Correcting Czech Case

The new formemes now distinguish adjectival and nominal usage of cardinal numerals (cf. also Section 5.2), e.g. the number in *5 potatoes* is now assigned the `adj:attr` formeme, whereas *Apollo 11* is given `n:attr`. The new situation is analogous in Czech, with nominal usages of numerals having their morphological case marked in formemes.

To reduce data sparsity in the new formemes version, we counter the inconsistent syntactic behavior of Czech cardinal numerals, where 1-4 behave like adjectives but other numerals behave like nouns and shift their semantically governing noun to the position of a genitive attribute. An example of this change is given in Fig. 2.



The word *banán* is in genitive (`n:2`), but would have an accusative (`n:4`) form if the numeral behaved like an adjective.

Figure 2: Case correction with numerals in Czech.

5.5 Adjectives: Nominal Usage and Case

The new formemes address the usage of adjectives in the syntactic position of nouns (cf. Section 5.2), which occurs only rarely, thus preventing sparse values, namely in these syntactic positions:

- *The subject.* We replaced the originally assigned `adj:compl` value, which was impossible to tell from adjectival objects, with the formeme a noun would have in the same position, e.g. in the sentence *Many of them were late*, the subject *many* is assigned `n:subj`.
- *Prepositional phrases.* Syntactic behavior of adjectives is identical to nouns here; we thus assign them the formeme values a noun would receive in the same position, e.g. `n:of+X` instead of `adj:of+X` in *He is one of the best at school.*

In Czech, we detect nominal usage of adjectives in verbal direct objects as well, employing large-coverage valency lexicons (Lopatková et al., 2008; Hajič et al., 2003).

Instead of assigning the `compl` value in Czech, our formemes revision includes the case of adjectival complements, which depends on the valency of the respective verb.

5.6 Mutual Information Across Languages

The changes described above have been motivated not only by theoretical linguistic description of the languages in question, but also by the intended usage within the TectoMT translation system. Instead

of retraining the translation model after each change, we devised a simpler and faster estimate to measure the asset of our innovations: using Mutual Information (MI) (Manning and Schütze, 1999, p. 66) of formemes in Czech and English trees.

We expect that an inter-language MI increase will lead to lower noise in formeme-to-formeme translation dictionary (Bojar et al., 2009, cf. Section 3.2), thus achieving higher MT output quality.

Using the analysis pipeline from CzEng1.0, we measured the inter-language MI on sentences from the Prague Czech-English Dependency Treebank (PCEDT) 2.0 (Bojar et al., 2012a). The overall results show an MI increase from 1.598 to 1.687 (Bojar et al., 2012b). Several proposed markup changes have been discarded as they led to an inter-language MI drop; e.g. removing the `v:rc` relative clause formeme or merging the `v:attr` and `adj:attr` values in English.

6 Experimental Usage

We list here our experiments with the newly developed annotation: an NLG experiment aimed at assessing the impact of formemes on the synthesis phase of the TectoMT system, and the usage in the English-Czech MT as a whole.

6.1 Czech Synthesis

The synthesis phase of the TectoMT system relies heavily on the information included in formemes, as its rule-based blocks use solely formemes and grammar rules to gradually change a deep tree node into a surface subtree.

To directly measure the suitability of our changes for the synthesis stage of the TectoMT system, we used a Czech-to-Czech round trip—deep analysis of Czech PDT 2.0 development set sentences using the CzEng 1.0 pipeline (Bojar et al., 2012b), followed directly by the synthesis part of the TectoMT system. The results were evaluated using the BLEU metric (Papineni et al., 2002) with the original sentences as reference; they indicate a higher suitability of the new formemes for deep Czech synthesis (see Table 2).

6.2 English-Czech Machine Translation

To measure the influence of the presented formeme revision on the translation quality, we compared two translation scenarios—one using the original formemes and the second using the revised formemes in the formeme-to-formeme translation model. Due to time reasons, we were able to train both translation models only on $1/2$ of the CzEng 1.0 training data.

Version	BLEU
Original formemes	0.6818
Revised formemes	0.7092

Table 2: A comparison of formeme versions in Czech-to-Czech round trip.

Version	BLEU
Original formemes	0.1190
Revised formemes	0.1199

Table 3: A comparison of formeme versions in English-to-Czech TectoMT translation on the WMT12 test set.

The results in Table 3 demonstrate a slight[6] BLEU gain when using the revised formemes version. The gain is expected to be greater if several rule-based modules of the transfer phase are adapted to the revisions.

7 Conclusion and Further Work

We have presented a systematic and theoretically supported revision of a surface morpho-syntactic markup within a deep dependency annotation scenario, designed to facilitate the TectoMT transfer phase. Our first practical experiments proved the merits of our innovations in the tasks of Czech synthesis and deep structural MT as a whole. We have also experimented with formemes in the functor assignment (semantic role labelling) task and gained moderate improvements (ca. 1-1.5% accuracy).

In future, we intend to tune the rule-based parts of our MT transfer for the new version of formemes and examine further possibilities of data sparsity reduction (e.g. by merging synonymous formemes). We are also planning to create formeme annotation modules for further languages to widen the range of language pairs used in the TectoMT system.

[6] Significant at 90% level using pairwise bootstrap resampling test (Koehn, 2004).

References

O. Bojar and Z. Žabokrtský. 2009. CzEng 0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92.

O. Bojar, D. Mareček, V. Novák, M. Popel, J. Ptáček, J. Rouš, and Z. Žabokrtský. 2009. English-Czech MT in 2008. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 125–129. Association for Computational Linguistics.

O. Bojar, J. Hajič, E. Hajičová, J. Panevová, P. Sgall, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas, J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová, and Z. Žabokrtský. 2012a. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, May. ELRA, European Language Resources Association. In print.

O. Bojar, Z. Žabokrtský, O. Dušek, P. Galuščáková, M. Majliš, D. Mareček, J. Maršík, M. Novák, M. Popel, and A. Tamchyna. 2012b. The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, May. ELRA, European Language Resources Association. In print.

J. Hajič, J. Panevová, Z. Urešová, A. Bémová, V. Kolárová, and P. Pajas. 2003. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9, pages 57–68.

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

G. Helbig and W. Schenkel. 1969. *Wörterbuch zur Valenz und Distribution deutscher Verben*. VEB Bibliographisches Institut, Leipzig.

P. Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, Barcelona, Spain.

M. Lopatková, Z. Žabokrtský, V. Kettnerová, and K. Skwarska. 2008. *Valenční slovník českých sloves*. Karolinum, Prague.

C.D. Manning and H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press.

D. Marcu, W. Wang, A. Echihabi, and K. Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 44–52. Association for Computational Linguistics.

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 Metrics (MATR)*, pages 201–206. Association for Computational Linguistics.

A. Menezes and S. D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the workshop on Data-driven methods in machine translation - Volume 14*, DMMT '01, pages 1–8, Stroudsburg, PA. Association for Computational Linguistics.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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

M. Popel, D. Mareček, N. Green, and Z. Žabokrtský. 2011. Influence of parser choice on dependency-based MT. In Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 433–439, Edinburgh, UK. Association for Computational Linguistics.

J. Ptáček and Z. Žabokrtský. 2006. Synthesis of Czech sentences from tectogrammatical trees. In *Text, Speech and Dialogue*, pages 221–228. Springer.

C. Quirk, A. Menezes, and C. Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 271–279. Association for Computational Linguistics.

M. Razímová and Z. Žabokrtský. 2006. Annotation of grammatemes in the Prague Dependency Treebank 2.0. In *Proceedings of the LREC 2006 Workshop on Annotation Science*, pages 12–19.

M. Ševčíková-Razímová and Z. Žabokrtský. 2006. Systematic parameterized description of pro-forms in the Prague Dependency Treebank 2.0. In J. Hajič and J. Nivre, editors, *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT)*, pages 175–186, Prague.

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

P. Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague.

Z. Urešová. 2009. Building the PDT-VALLEX valency lexicon. In *On-line proceedings of the fifth Corpus Linguistics Conference*. University of Liverpool.

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*, StatMT '08, pages 167–170, Stroudsburg, PA. Association for Computational Linguistics.

Z. Žabokrtský. 2005. *Valency Lexicon of Czech Verbs*. Ph.D. thesis, Charles University in Prague.

Z. Žabokrtský. 2010. *From Treebanking to Machine Translation*. Habilitation thesis, Charles University in Prague.

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, Suntec, Singapore.