

Czechizator – Čechizátor

Rudolf Rosa

Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics,
Malostranské náměstí 25, 118 00 Prague, Czech Republic
rosa@ufal.mff.cuni.cz

Abstract: We present a lexicon-less rule-based machine translation system from English to Czech, based on a very limited amount of transformation rules. Its core is a novel translation module, implemented as a component of the TectoMT translation system, and depends massively on the extensive pipeline of linguistic preprocessing and post-processing within TectoMT. Its scope is naturally limited, but for specific texts, e.g. from the scientific or marketing domain, it occasionally produces sensible results.

Prezentujeme lexikon-lesový rule-bazovaný systém machín translace od Engliše Čecha, který bazoval na verově limitované amountu rulů transformace. Jeho kor je novelový modul translace, implementovalo jako komponent systému translace tektomtu a dependuje masivně na extensivní pipelínu lingvistické preprocesování a postprocesovat v Tektomtu. Jeho skop je naturálně limitovaná, ale pro specifické texty z například scientifické nebo marketování doménu okasionálně produkuje sensibilní resulty.

1 Introduction and Motivation

In this work, we present Czechizator, a lexicon-less rule-based machine translation system from English to Czech.

Lexicon-less approach to machine translation has already been successfully applied to closely related languages – e.g. the Czech-Slovak machine translation system Česílko [3, 4] featured a rule-based lexicon-less transformation component for handling OOV (out-of-vocabulary) words. For transliteration, which can be thought of as a low-level translation, rule-based systems are also common. However, in this work, we decided to tackle a harder problem: to use a similar approach for a full translation between a pair of only weakly related languages, namely English and Czech.

While we believe that it is impossible to achieve high-quality or even reasonable-quality general-domain translation without a large lexicon, we attempt to investigate to what degree this is possible if the domain is somewhat special. Specifically, we target the domain of scientific texts (or, more precisely, abstracts of scientific papers), which contain a large amount of terms that tend to be rather similar even across more distant languages. In this way, we operate on a pair of languages which are typologically different but lexically close. Moreover, we crucially rely on the strong linguistic abstractions provided by the TectoMT machine translation system [15], which boasts to operate

on a deep layer of language representation where typological differences of languages become quite transparent, as the meaning itself, rather than the form, is captured. Abstracting away from both lexical and typological differences in this way, a smallish set of rules and heuristics should be sufficient to obtain a competitive machine translation system.

While the main focus of our work is to test the degree to which the aforementioned hypothesis is valid, our work has practical implications as well. The number of terms used in scientific texts is enormous, many of them being rare in parallel corpora or even newly created and thus bound to constitute OOV items for machine translation systems. However, as there seems to be some regularity in the way that English terms are adapted in Czech, it should be possible to use a lexicon-less system as an additional component in a standard machine translation system to handle OOVs. It may also be beneficial in scenarios where low-quality but light-weight translation system is preferred over a full-fledged but resource-heavy system.¹

Another use-case is machine-aided translation of scientific paper abstracts, as the Czechizator output should often be a good starting point for creating the final translation by post-editing.

Before explaining the approach we used to implement the translation model, we present a set of three sample outputs of Czechizator, applied to abstracts of two scientific papers (Table 1, Table 2), and one marketing text,² (Table 3). Also, as an additional example, the abstract of this paper is provided both in English and in its Czechization.

2 Approach

2.1 TectoMT

TectoMT [15, 1] is a highly modular linguistically oriented machine translation system, featuring a deep-linguistic three-step processing pipeline of analysis, transfer, and

¹However, TectoMT itself is rather resource-heavy even when the lexical models are omitted, so even though the component that we implemented is very light-weight, the complete system that it relies on is not – using the Czechizator model instead of the base models in TectoMT only brings a 15% speedup and 40% RAM cut, which is probably not worth the quality drop in any realistic scenario.

²The text was obtained from <https://www.accenture.com/cz-en/strategy-index>

Source	Czechization	Reference translation
Chimera is a machine translation system that combines the TectoMT deep-linguistic core with Moses phrase-based MT system. For English–Czech pair it also uses the Depfix post-correction system. All the components run on Unix/Linux platform and are open source (available from CPAN Perl repository and the LINDAT/CLARIN repository). The main website is https://ufal.mff.cuni.cz/tectomt . The development is currently supported by the QTLearn 7th FP project (http://qtlearn.eu).	Chimera je systém machín translace, který kombinuje díp-lingvistické kor tek-tomtu z fraze-bazovaného MT systému mozesu. Pro Engliše – čechová pér také uzuje systém post-korekce Dep-fix. Všechny komponenty runují v Unix / platformu Linuxu a jsou open-ová sourc (avélabilní z CPAN Perla repositorie a LINDAT / CLARIN repos-itorie). Hlavní webová stránka je https://ufal.mff.cuni.cz/tectomt . Devel-opment kurentně je suport FP projektem 7th qtlípu (http://qtlearn.eu).	Chimera systém strojového překlada, který kombinuje hluboce lingvistické jádro TectoMT s frázovým strojovým překladačem Moses. Pro anglicko-český překlad také používá post-editovací systém Depfix. Všechny komponenty běží na platformě Unix/Linux a jsou open-source (dostupné z Perlového repositáře CPAN a repositáře LIN-DAT/CLARIN). Hlavní webová stránka je https://ufal.mff.cuni.cz/tectomt . Vývoj je momentálně podporován projektem QTLearn ze 7th FP (http://qtlearn.eu).

Table 1: Abstract of a scientific paper [7], its Czechization, and a reference translation by its author.

Source	Czechization
We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.	Propozujeme 2 novelová architektury modelů, že komputují kontinuální reprezentace vektorů vordů od verově largových setů dat. Kvalita těchto reprezentací je mísur ve vord similarita tasku a resulty jsou kompar s previálně nejgúdošími, perfor-mují, techniky, kteří bazovali na diferentových typech neurál-ních networků. Observujeme largové improvementy akurace v muchově lovovší komputacionální kosti, tj. takuje méně než Daie, aby se lírnovalo hajové vektory vordu kvality z dat vordů 1.6 bilionu, která setovala. Furtermorově šovujeme, že tyto vek-tory providují state-of-te-artovou performance na našem testu, který setoval, že mísurují syntaktické a semantické vord simi-larity.

Table 2: Abstract of a scientific paper [6] and its Czechization.

Source	Czechization
Accenture Operations combines technology that digitizes and automates business processes, unlocks actionable insights, and delivers everything-as-a-service with our team’s deep industry, functional and technical expertise. So you can confidently chart your course to consuming your core business services on demand, accelerate innovation and speed to market. Welcome to the "as-a-service" business revolution. Accenture Strategy shapes our clients’ future, combining deep business insight with the understanding of how technology will impact industry and business models. Our focus on issues related to digital disruption, redefining competitiveness, operating and business models as well as the workforce of the future helps our clients find future value and growth in a digital world. Whether focused on strategies for business, technology or operations, Accenture Strategy has the people, skills and experience to effectively shape client value. We offer highly objective points of view on C-suite themes, with an emphasis on business and technology, leveraging our deep industry experience. That’s high performance, delivered.	Operacions acenturu kombinuje technologii, která digitizuje a automuje procesy businosti, unlokuje akcionabilní insajty a deliveruje everyting-as-a-servicová s funkcionální a technickou expertizou dípové industrie našeho tímu. Tak konfidentně můžete chartovat svůj kours, konsumuje vaše service businosti kor na demandu, aceleratové inovaci a spídu marketu. Velkomujte „as-a-service“ revoluce businosti. Strategie acenturu šapuje futur našich klientů, kombinuje dípovou insajt businosti s understandováním, jak technologie impaktuje a industrie businosti modely. Náš fokus na isu, kteří relovali s digitálním disruptcí, kteří redefinují kompetitivnost, operatování a businost modely, i vorkforc futur helpuje, naši klienti findují futurovou valu a grovt v digitální vorldu. Vhetr fokusoval na strategie pro businost, technologie nebo operací strategii acenturu, má peoply, skily a experience, aby efek-tivně šapovali valu klienta. Oferujeme hajně objektivní pointy vievu na k-suitových temech s emfasi na businost a technologii, leveraguje naši dípovou experience industrie. Které je hajová performanc, který deliveroval.

Table 3: A marketing text from Accenture.com and its Czechization.

synthesis. TectoMT is implemented in Treex [8, 13], using a representation of language based on the Functional Generative Description [11].

The first step in the translation pipeline is to perform a linguistic analysis of each source (input) sentence up to t-layer, obtaining a deep-syntactic representation of the sentence (t-tree). On t-layer, each full (autosemantic) word is represented by a t-node with a t-lemma and a set of linguistic t-attributes (such as functor, formeme, number, gender, deep tense) that capture the function of the word. Inflections and auxiliary words are not explicitly represented, but their functions are captured by the attributes of the t-nodes.

Each source t-tree is then isomorphically transferred to a target t-tree. In the standard TectoMT setup, the t-lemma of each t-node is translated by models that have been trained on large parallel data. The other t-attributes are then transferred by a pipeline featuring both rule-based and machine-learned steps.

Finally, the target sentence is synthesized from the t-tree. This step relies heavily on a morphological generator [12], which is able to generate a word form based on the word lemma and a set of morphological feature values. For the highly fleective Czech language, this is a challenging task; even though we employ a state-of-the-art generator, it is sometimes unable to generate the requested word form, especially when the lemma is unknown to the generator.

TectoMT can (and does by default) use a weighted interpolation of multiple translation models to generate translation candidates [10]. This makes it easy to replace or complement the existing models with new models, such as our Czechizator model.

2.2 Czechizator translation model

The Czechizator translation model attempts to Czechize each English t-lemma, unless it is marked as a named entity. To Czechize the lemma, it applies the following resources, which we manually constructed:

- a shortlist of 36 lemma translations, focusing on words that we believe to be auxiliaries rather than full words (and thus presumably should be dropped by the t-analysis and represented by t-attributes, but in fact constitute t-lemmas),³ and on cardinal numbers (which presumably should be converted to a language-independent representation by TectoMT analysis, but are not),
- a set of 43 transformation rules based on semantic part of speech of the t-node and the ending of its t-lemma (noun rules are provided as an example in Table 4), and
- a transliteration table, consisting of 33 transliteration rules.⁴

English ending	Czechized ending
-sion	-se
-tion	-ce
-ison	-ace
-ness	-nost
-ise	-iza
-ize	-iza
-em	-ém
-er	-r
-ty	-ta
-is	-e
-in	-ín
-ine	-ín
-ing	-ování
-cy	-ce
-y	-ie

Table 4: A list of ending-based transformations of noun lemmas.

The transformations are generally applied sequentially, but forking is possible at some places, and so multiple alternative Czechizations may be generated; TectoMT uses a Hidden Markov Tree Model [14] (instead of a language model) to eventually select the best combination of t-lemmas (and other t-attributes). However, as the Czechizations are usually OOVs for the HMTM, typically the first candidate gets selected. The target semantic part-of-speech identifier is also generated, based on the source semantic part-of-speech and the t-lemma ending; this is important for the subsequent synthesis steps.

It should be noted that the current implementation of Czechizator is rather a proof-of-concept than an attempt on a professional translation model. If one was to follow this research path in future, it would be presumably more appropriate to learn the regular transformations from parallel (or comparable) corpora, extracting pairs of similar words that are translations of each other and generalizing the transformation necessary to convert one into the other, as well as learning to identify the cases in which a transformation should be applied. Similar methods could be used as were applied e.g. in the semi-supervised morphological generator Flect [2].

Czechizator uses the standard TectoMT translation model interface, and can thus be easily and seamlessly plugged into the standard TectoMT pipeline, either replacing or complementing the base lexical translation models.

2.3 Surrogate lemma inflection

As Czechizator generates many weird and/or non-existent lemmas, it is an expected consequence that the morphological generator is often unable to inflect these lemmas. For

³be, have, do, and, or, but, therefore, that, who, which, what, why, how, each, other, then, also, so, as, all, this, these, many, only, main, mainly

⁴As an example, we list several of the transliteration rules here: th→t, ti→ci, ck→k, ph→f, sh→š, ch→ch, cz→č, qu→kv, igh→aj, gh→ch, gu→gv, dg→dž, w→v, c→k.

Ending	Surrogate lemma
-ovat	kupovat
-ání	plavání
-í	jarní
-ý	mladý
-o	město
-e	růže
-a	žena
-ost	kost
-ě	mladě
-h, k, r, d, t, n, b, f, l, m, p, s, v, z	svrab
-ž, š, ř, č, c, j, d', t', ň	muž

Table 5: List of surrogate lemmas for given endings. The matched ending gets deleted from the target lemma, obtaining the target pseudo-stem, except for the last two cases (matching hard or soft final consonants), where even the final consonant is part of the stem.

this reason, we enriched the word form generation component of TectoMT⁵ with a last-resort inflection step.⁶ If the morphogenerator is unable to generate the inflection, we use a set of simple ending-based rules to find a *surrogate lemma*, as listed in Table 5,⁷ inflect the surrogate lemma, strip its ending, and apply it to the target lemma. We focus on endings generated by the Czechizator translation module, but we aimed for high coverage, and successfully managed to employ the last-resort inflector even into the base TectoMT translation.

For example, if one is to inflect the pseudo-adjective “largový” (Czechization of “large”) for the feminine accusative, we replace it with the surrogate lemma (“mladý”) that corresponds to its ending (“-ý”), obtain its feminine accusative inflection from the morphogenerator (“mladou”), strip the matched ending from both of the lemmas, obtaining pseudo-stems (“largov”, “mlad”), strip the surrogate pseudo-stem (“mlad”) from the surrogate inflection (“mladou”) to obtain the inflection ending (“-ou”), and join the ending with the target pseudo-stem (“largov”) to obtain the target inflection (“largovou”).

3 Evaluation

3.1 Dataset

To automatically evaluate the translation quality by standard methods, we collected a small dataset, consisting of Czech and English abstracts of scientific papers. Specifically, we collected the abstracts of papers of authors from

⁵<https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2A/CS/GenerateWordforms.pm>

⁶<https://github.com/ufal/treex/commit/363d1b18f7140e0cb687ed8deebc4ac4a1051080>

⁷Although there exists a set of commonly used lemmas to represent the basic Czech paradigms, we sometimes use a different lemma – to avoid unnecessary ambiguity, and to simplify the application of the ending to the target lemma (we avoid surrogate lemmas that exhibit changes on the root during inflection).

Setup	BLEU	NIST
Untranslated source	3.41	1.13
No model	2.85	1.62
Czechizator	3.01	2.08
Base TectoMT	8.75	3.62
Base + Czechizator	8.33	3.57

Table 6: Automatic evaluation scores on the ÚFAL abstracts dataset[9].

the Institute of Formal and Applied Linguistics at Charles University in Prague, who are obliged to provide both a Czech and an English abstract for each of their publications. These are then stored in the institute’s database of publications, Biblio,⁸ and can be accessed through a regularly generated XML dump.⁹

The collected parallel corpus, aligned on the document level, e.g. on individual abstracts, contains 1,556 pairs of abstracts, totalling 121,386 words on the English side and 76,812 words on the Czech side.¹⁰ We did not perform any filtering of the data, apart from filtering out incomplete entries (missing the Czech or the English abstract) and replacing newlines and tabulators by spaces (solely for technical reasons). The dataset is publicly available [9].

3.2 Evaluation and discussion

Automatic evaluation with BLEU and NIST was performed with the MTrics tool [5]. We evaluated several candidate translations: the untranslated English source texts, TectoMT with no lexical model, TectoMT with the Czechizator model, TectoMT with an interpolation of its base lexical models (the default setup of TectoMT), and TectoMT with an interpolation of Czechizator and the base lexical models.

While translation quality of the Czechizator outputs is clearly well below the base TectoMT system, the results show that Czechizator does manage to produce some useful output – its scores are significantly higher than that of TectoMT with no lexical translation model. This shows that lexicon-less translation is somewhat possible in our setting, although on average it is far from competitive – at least with the current version of Czechizator, which is a rather basic proof-of-concept implementation, lacking numerous simple and obvious improvements that could easily be performed and would presumably lead to further significant increases of translation quality. However, as with many rule-based systems for natural language processing, the code complexity and especially the amount of manual tuning necessary to push the performance further and further is likely to grow very quickly.

⁸<http://ufal.mff.cuni.cz/biblio/>

⁹<https://svn.ms.mff.cuni.cz/trac/biblio/browser/trunk/xmldump>

¹⁰The difference in the sizes is partially caused by the fact that usually, the English abstract is the full original, and its Czech translation is often shortened considerably by the authors.

Manual inspection of the outputs (see also the examples in the beginning of this paper) showed that the chosen domain is quite suitable for lexicon-less translation, but the proportion of autosemantic words that cannot be simply transformed from English to Czech without a lexicon is still rather high – high enough to make many of the sentences barely comprehensible. We therefore acknowledge that at least a small lexicon would be necessary to obtain reasonable translations for most sentences. On the other hand, we observed many phrases, and occasionally even whole sentences, whose Czechizations were of a rather high quality and understandable to Czech speakers with minor or no difficulties. We thus find our approach interesting and potentially promising, although we believe that the amount of work needed to bring the system to a competitive level of translation quality would be by several orders of magnitude larger than that spent on creating the current system (which took less than one person-week). Still, we expect that for the given domain, developing such a rule-based system would constitute many times less work than building an open-domain system.

Thanks to the deep analysis and generation provided by TectoMT, the Czechizations tend to be rather grammatical, with words correctly inflected, even if non-sensical. Unfortunately, even grammatical errors occur rather frequently – some words are not inflected at all, some violate morphological agreement (e.g. in gender, case or number), etc. This can be explained by realizing that the complex TectoMT pipeline consists of many subcomponents, each operating with a certain precision, occasionally producing erroneous analyses. The most crucial stage seems to be syntactic parsing, which has been reported to have only approximately 85% accuracy, i.e. roughly 15% of dependency relations are assigned incorrectly; these typically manifest themselves as agreement errors in the Czechization output.

Evaluation of the main potential use case of Czechizator, i.e. complementing base TectoMT translation models for OOVs (Base + Czechizator setup), brought mixed results. There is a small deterioration in the automatic scores, and subsequent manual inspection showed that Czechizator can target OOVs only semi-successfully. It can offer a Czechization of any OOV term, which is often correct (e.g. “anafora” for English “anaphora”, “interlingvální” for “interlingual”, “hypotaktický” for “hypotactical”, or “cirkumfixální” for “circumfixal”), but sometimes the Czechization is not correct (e.g. “businost” for “business”, “hands-onový” for “hands-on”, or “kolokaty” for “collocations”). In many cases, a Czechization of the term is simply not used in practice, and is less understandable to the reader than the original English form (e.g. “kejnotový” for “keynote”, “veb-pagová” for “web-page”, “part-of-spích” for “part-of-speech”, or “kros-langvaž” for “cross-language”). Czechizator also often generates a form that is plausible but rarely or never used, although one may think that the Czechized form may become the standard Czech translation in future, and is mostly under-

standable to readers (e.g. “tríbank” for “treebank”, “tvít” for “tweet”, or “kros-lingvální” for “cross-lingual” – here the base models generated a rather nonsensical “lingual kříže”). Unfortunately, it also often Czechizes named entities, even though we explicitly avoid them if they are marked by the analysis; this seems to be primarily a short-coming (or unsuitability for this task) of the named entity recognizer used [12], which seems to favour precision over recall. Still, Czechizator can sometimes provide a better translation than the base models, even in cases where the term is not an OOV – such as the word “post-editing”, which the base models translate into a confusing “poúprava”, while Czechizator provides an acceptable translation “post-editování”.¹¹

In general, we believe that, if appropriate attention is paid to the identified issues, such as named entities avoidance, Czechizator has the potential of usefully complementing the base TectoMT translation models, especially in handling OOV terms.

4 Conclusion

We implemented a rule-based lexicon-less English-Czech translation model into TectoMT, called Czechizator. The model is based on a set of simple rules, mainly following regularities in adoption of English terms into Czech. Czechizator has been especially designed for and applied to the domain of abstracts of scientific papers, but also provides interesting results for texts from the marketing domain.

We automatically evaluated Czechizator on a collection of abstracts of computational linguistics papers, showing inferior but promising results in comparison with the base TectoMT models; the highest observed potential is in employing Czechizator as an additional TectoMT translation model for out-of-vocabulary items.

Czechizator is released as an open-source Treex module in the main Treex repository on Github,¹² and is also made available as an online demo.¹³

Acknowledgments

This research was supported by the grants GAUK 1572314, and SVV 260 333. This work has been using language resources and tools developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

¹¹Other such examples include the Czechization “reimplementace” for “reimplementation” instead of “znovuprovádění”, or “post-nominální” for “post-nominal” instead of “pojmenovitý”.

¹²<https://github.com/ufal/treex/blob/master/lib/Treex/Tool/TranslationModel/Rulebased/Model.pm>

¹³<http://ufallab.ms.mff.cuni.cz/~rosa/czechizator/input.php>

References

- [1] Ondřej Dušek, Luís Gomes, Michal Novák, Martin Popel, and Rudolf Rosa. New language pairs in tectoMT. In *Proceedings of the 10th Workshop on Machine Translation*, pages 98–104, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics, Association for Computational Linguistics.
- [2] Ondřej Dušek and Filip Jurčiček. 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 (Volume 1: Long Papers)*, pages 451–461, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics, Association for Computational Linguistics.
- [3] Jan Hajič, Vladislav Kuboň, and Jan Hric. Česílko - an MT system for closely related languages. In *ACL2000, Tutorial Abstracts and Demonstration Notes*, pages 7–8. ACL, ISBN 1-55860-730-7, 2000.
- [4] Petr Homola and Vladislav Kuboň. Česílko 2.0, 2008.
- [5] Kamil Kos. Adaptation of new machine translation metrics for Czech. Bachelor’s thesis, Charles University in Prague, 2008.
- [6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [7] Martin Popel, Roman Sudarikov, Ondřej Bojar, Rudolf Rosa, and Jan Hajič. TectoMT – a deep-linguistic core of the combined chimera MT system. *Baltic Journal of Modern Computing*, 4(2):377–377, 2016.
- [8] Martin Popel and Zdeněk Žabokrtský. TectoMT: Modular NLP framework. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrun Helgadóttir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg, 2010. Iceland Centre for Language Technology (ICLT), Springer.
- [9] Rudolf Rosa. Czech and English abstracts of ÚFAL papers, 2016. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.
- [10] Rudolf Rosa, Ondřej Dušek, Michal Novák, and Martin Popel. Translation model interpolation for domain adaptation in TectoMT. In Jan Hajič and António Branco, editors, *Proceedings of the 1st Deep Machine Translation Workshop*, pages 89–96, Praha, Czechia, 2015. ÚFAL MFF UK, ÚFAL MFF UK.
- [11] Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer, 1986.
- [12] Jana Straková, Milan Straka, and Jan Hajič. 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, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- [13] Zdeněk Žabokrtský. Treex – an open-source framework for natural language processing. In Markéta Lopatková, editor, *ITAT*, volume 788, pages 7–14, Košice, Slovakia, 2011. Univerzita Pavla Jozefa Šafárika v Košiciach.
- [14] Zdeněk Žabokrtský and Martin Popel. Hidden Markov tree model in dependency-based machine translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 145–148, Suntec, Singapore, 2009. Association for Computational Linguistics.
- [15] Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogramatics used as transfer layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, OH, USA, 2008. Association for Computational Linguistics.