PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 102 OCTOBER 2014 47–56

Depfix, a Tool for Automatic Rule-based Post-editing of SMT

Rudolf Rosa

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

We present Depfix, an open-source system for automatic post-editing of phrase-based machine translation outputs. Depfix employs a range of natural language processing tools to obtain analyses of the input sentences, and uses a set of rules to correct common or serious errors in machine translation outputs. Depfix is currently implemented only for English-to-Czech translation direction, but extending it to other languages is planned.

1. Introduction

Depfix is an automatic post-editing system, designed for correcting errors in outputs of English-to-Czech statistical machine translation (SMT) systems. An approach based on similar ideas was first used by Stymne and Ahrenberg (2010) for English-to-Swedish. Depfix was introduced in (Mareček et al., 2011), and subsequent improvements were described especially in (Rosa et al., 2012b) and (Rosa et al., 2013). For a comprehensive description of the whole Depfix system, please refer to (Rosa, 2013). An independent implementation for English-to-Persian exists, called Grafix (Mohaghegh et al., 2013).

Depfix consists of a set of rule-based fixes, and a statistical component.[1] It utilizes a range of NLP tools, especially for linguistic analysis of the input (taggers, parsers, named entity recognizers…), and for generation of the output (morphological generator, detokenizer…). Depfix operates by analyzing the input sentence, and invoking a pipeline of error detection and correction rules (called *fixes*) on it.

Depfix is one of the components of Chimera (Bojar et al., 2013b; Tamchyna et al., 2014), the current state-of-the-art system for English-to-Czech machine translation

[1] However, the vital part of Depfix are the rule-based fixes.

 Corresponding author: rosa@ufal.mff.cuni.cz
Cite as: Rudolf Rosa. Depfix, a Tool for Automatic Rule-based Post-editing of SMT. The Prague Bulletin of Mathematical Linguistics No. 102, 2014, pp. 47–56. doi: 10.2478/pralin-2014-0013.

System	WMT 2011	WMT 2012	System	WMT 2011	WMT 2012
CU Bojar	+**0.47**	+0.07	JHU	+**0.42**	+**0.32**
CU Tamchyna	+**0.46**	+0.02	SFU	–	+**0.41**
CU TectoMT	−**0.10**	−0.02	EuroTran	+**0.21**	+**0.15**
CU Zeman	+**0.73**	+**0.34**	Microsoft Bing	–	+**0.37**
UEDIN	+**0.64**	+**0.23**	Google Translate	+**0.23**	0.00

Table 1. Automatic evaluation of the Depfix system. Adapted from (Rosa, 2013). Change of BLEU score when Depfix was applied to the output of the system is reported. Statistically significant results are marked by bold font.

(MT) – the other components are TectoMT (Žabokrtský et al., 2008) and factored Moses (Koehn et al., 2007). Chimera has ranked as the best English-to-Czech MT system in the last two translation tasks of the Workshop on Statistical Machine Translation (WMT) (Bojar et al., 2013a, 2014). Depfix is currently being developed in the frame of QTLeap,[2] a project focusing on quality translation by deep language engineering approaches.

Depfix is a stand-alone system, and can be used to post-process outputs of any MT system. It particularly focuses on errors common in phrase-based SMT outputs; some of its components have been tuned using outputs of Moses. In a throughout evaluation on outputs of all systems participating in WMT in 2011 and 2012 (Callison-Burch et al., 2011, 2012), applying Depfix led to a statistically significant improvement in BLEU score in most cases, as shown in Table 1.[3]

Depfix is implemented in the Treex framework (Popel and Žabokrtský, 2010).[4] Instructions on obtaining and using Depfix can be found on `http://ufal.mff.cuni.cz/depfix`. We release Depfix under the GNU General Public License v2 to encourage its improvement and adaptation for other languages, as well as to serve as inspiration for other researchers.

2. Tools

Depfix basically operates by observing and modifying morphological tags. Therefore, the two following tools are vital for operation of Depfix:

- A **lemmatizing tagger** (or a tagger and a lemmatizer) is an analysis tool that assigns the word form of each token in the sentence with a combination of lemma

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

[3] We did not perform this kind of evaluation in the following years of WMT, as we focused on the Chimera hybrid system instead.

[4] `http://ufal.mff.cuni.cz/treex`

and tag. We use MorphoDiTa (Straková et al., 2014) for Czech and Morče (Spoustová et al., 2007) for English.
- A **morphological generator** is a generation tool inverse to the tagger: for a given combination of lemma and tag, it generates the corresponding word form. We use Hajič's Czech morphological generator (Hajič, 2004).

For Czech, we use the Prague dependency treebank positional tagset (Hajič, 1998), which marks 13 morphological categories, such as part-of-speech, gender, number, case, person, or tense. One of the properties of this tagset, which is very useful for us, is that the lemmas and morphological categories are fully disambiguated – for a given combination of lemma and tag, there is at most one corresponding word form (the opposite does not hold, as many Czech paradigms have the same word repeated several times).

For English, we use the Penn treebank tagset (Marcus et al., 1993), which marks only few morphological categories, such as singular/plural number for nouns, but does not distinguish e.g. verb person (except for 3rd person singular in present simple tense).

Apart from the tagger and the morphological generator, many other tools are used in Depfix. We currently use the following, which are either implemented within the Treex framework, or are external tools with Treex wrappers:
- a rule-based Treex tokenizer and detokenizer
- a word aligner – GIZA++ (Och and Ney, 2003)
- a dependency parser – MST parser for English (McDonald et al., 2005), and its variations for Czech: a version by Novák and Žabokrtský (2007) adapted for Czech in the basic version of Depfix, or MSTperl by Rosa et al. (2012a) adapted for SMT outputs in full Depfix
- a dependency relations labeller (as the MST parser returns unlabelled parse trees) – a rule-based Treex labeller for English, and a statistical labeller by Rosa and Mareček (2012) for Czech
- a named entity recognizer – Stanford NER for English (Finkel et al., 2005), and a simple Treex NER for Czech
- a rule-based Treex converter to tectogrammatical (deep syntax) dependency trees

There are also other tools that we do not currently use (because they are not part of Treex yet, some of them probably do not even exist yet), but we believe that they would be useful for Depfix as well:
- a full-fledged named entity recognizer for Czech
- a coreference resolver
- a fine-grained tagger for English (that would mark e.g. verb person or noun gender)
- a well-performing labeller for tectogrammatical trees (the current Treex one is rather basic and lacks proper analysis of verbs, negation, etc., especially for English)

When porting Depfix to a new language, acquiring the NLP tools is a necessary first step. Unfortunately, in the Treex framework, support for languages other than Czech and English is currently very limited. However, one can make use of the HamleDT project (Zeman et al., 2012),[5] which collects dependency treebanks for various languages and harmonizes their tagsets and dependency annotation styles to a common scheme. We believe this to be an ideal resource for training a tagger as well as a dependency parser for any of the languages covered – HamleDT 2.0 currently features 30 treebanks and is still growing, and there are also plans of its tighter integration with Treex.

3. Fixing Rules

The main part of Depfix is a set of fixing rules (there are 28 of them in the current version). Most of the rules inspect the tag of a Czech word, usually comparing it to its source English counterpart and/or its Czech neighbours (usually its dependency parents or children), and if an error is spotted (such as incorrect morphological number – e.g. the Czech word is in singular but the source English word is in plural), the tag of the Czech word is changed to the correct one, and the morphological generator is invoked to generate the corresponding correct word form.[6]

Some of the rules also delete superfluous words (e.g. a subject pronoun that should be dropped), change word order (e.g. the noun modifier of a noun, which precedes the head noun in English but should follow it in Czech), or change tokenization and casing (by projecting it from the source English sentence where this seems appropriate).

The ideas for the rules are based on an analysis of errors in English-to-Czech SMT (Bojar, 2011), and the actual rules were implemented and tuned using the WMT 2010 test set (Callison-Burch et al., 2010) translated by Moses. For other language pairs, a similar error analysis, such as the Terra collection (Fishel et al., 2012), may be used as a starting point; however, the error analyses are typically not fine-grained enough to be used directly for Depfix rules implementation, and extensive manual tuning of the rules by inspecting SMT translation outputs is to be expected.

3.1. Example

Table 2 shows the operation of *FixPnom* rule.[7] In the sentence, there is an error in agreement of nominal predicate "zdrženliví" ("reticent$_{\mathtt{pl}}$") with the subject "Obama".

[5] `http://ufal.mff.cuni.cz/hamledt`

[6] It may happen that the new word form turns out to be identical to the original word form, as there are often many possible tags for a word – e.g. the nominative and accusative case of a noun is often identical. However, the fix may still be beneficial, as subsequent fixes might be helped by the corrected tag of the word – e.g. a fixed noun tag may induce a fix of a modifying adjective.

[7] The **`make`** `compare_log` Depfix command can be used to obtain this kind of information.

Source:	Obama has always been reticent in regards to his prize.
SMT output:	Obama byl vždy **zdrženliví** s ohledem na svou kořist.
Depfix output:	Obama byl vždy **zdrženlivý** s ohledem na svou kořist.
Fixlog:	Pnom: zdrženliví[AAM**P**1—-1A—-] zdrženlivý[AAM**S**1—-1A—-]

Table 2. Example of application of FixPnom *on a sentence from WMT10 dataset (translated by the CU-Bojar system)*

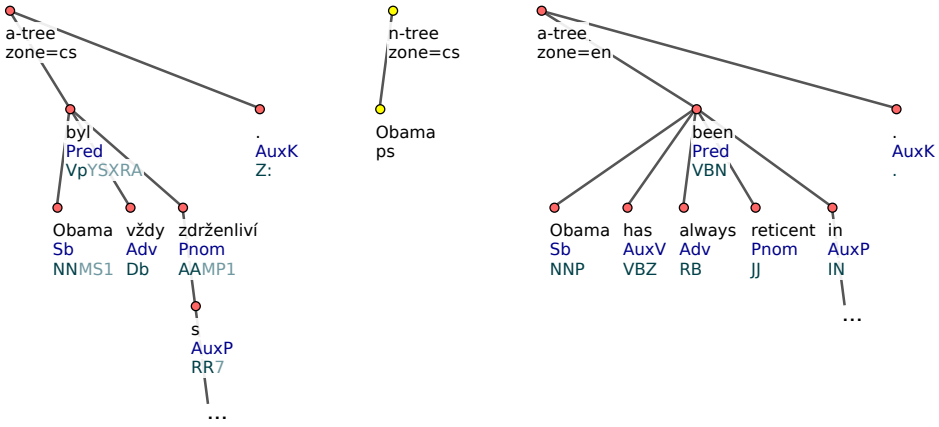

Figure 1. Part of the dependency parse tree of a Czech sentence before applying FixPnom*. Also showing the Czech named entity tree, and corresponding part of the source English dependency parse tree. Word forms, analytical functions, and morphological tags are shown.*

The morphological number (4th position of the tag) should be identical for both of the words, but it is not – it is singular ("S") for "Obama" but plural ("P") for "zdrženliví". See also Figure 1, which shows the parse tree of the Czech sentence (before applying the fix), the named entity tree for the Czech sentence, and the parse tree of the source English sentence.[8]

When the *FixPnom* rule is invoked on the word "zdrženliví", it realizes the following:

- the word is an adjective and its dependency parent is a copula verb, thus the word is a nominal predicate and the *FixPnom* rule applies here,

[8]These parse trees, as well as the tectogrammatical trees, are contained in intermediate `*.treex` files in the Depfix experiment directory, and can be viewed using Tree Editor TrEd – see `http://ufal.mff.cuni.cz/tred/`.

- there is a child of the parent verb ("Obama") which is marked as subject, and its English counterpart ("Obama") is also marked as subject, thus it should be in agreement with the nominal predicate,
- the subject is in singular, while the nominal predicate is in plural, thus the agreement is violated and should be fixed.

The rule therefore proceeds by fixing the error. This is done in two steps:

1. the tag of "zdrženliví" is changed by changing the number marker from "P" (plural) to "S" (singular), as indicated in the Fixlog in Table 2,[9]
2. the morphological generator is invoked to generate a word form that corresponds to the new tag; in this case, the word "zdrženlivý" is generated.

The *FixPnom* rule also checks and corrects agreement in morphological gender; however, agreement in gender is not violated in the example sentence.

4. Implementation

Depfix is implemented in the Treex framework, which is required to run it, and is operated from the command-line via Makefile targets. The commented source code of Depfix is in Perl and Bash. The fixing blocks are implemented as Treex blocks, usually taking a dependency edge as their input, checking it for the error that they fix, and fixing the child or parent node of the edge as appropriate.

The Depfix Manual (Rosa, 2014a), which provides instructions on installing and running Depfix, is available on the Depfix webpage. The installation consists of installing Treex and several other modules from CPAN, checking out the Treex subversion repository (which Depfix is contained in), downloading several model files, and making a test run of Depfix.

Depfix needs a Linux machine to run, with at least 3.5 GB RAM to run the basic version – i.e. without the MSTperl parser, which is adapted for SMT outputs (Rosa et al., 2012a; Rosa, 2014b), and without the statistical fixing component (Rosa et al., 2013). The full version, which achieves slightly higher BLEU improvement than the basic version, needs at least 20 GB to run.

Depfix takes source English text and its machine translation as its input, and provides the fixed translations as its output (all plain text files, one sentence per line). Processing a set of 3000 sentences by Depfix takes about 2 hours; processing a single sentence takes about 5 minutes (most of this time is spent by initializing the tools).[10]

[9] The fix is performed in this direction because the morphological number is more reliable with nouns in English-to-Czech translation, as noun number is explicitly marked in English while adjective number is not.

[10] These times are provided for illustration only, as they depend on the speed of the processor, the hard-drive, and other parameters of the machine.

5. Conclusion and Future Work

In this paper, we described Depfix, a successful automatic post-editing system system designed for performing rule-based correction of errors in English-to-Czech statistical machine translation outputs.

As a stand-alone tool, Depfix can be used to post-edit outputs of any machine translation system, although it focuses especially on shortcomings of the phrase-based ones, such as Moses. So far, we have implemented Depfix only for the English-to-Czech translation direction, although there exist similar systems for other languages by other authors. Depfix has been developed for several years, and is now a component of Chimera, the state-of-the-art machine translation system for English-to-Czech translation.

The future plans for Depfix development are directed towards extending it to new translation directions, starting with a refactoring to separate language-independent and language-specific parts, so that fixing rules for a new language pair can be implemented easily while reusing as much from the already implemented functionality as possible. Another future research path aims to complement or replace the manually written rules by machine learning techniques, with preliminary experiments indicating viability of such an approach.

To improve the ease of use of Depfix, we also wish to implement an online interface that would enable invoking Depfix remotely from the web browser or as a web service, with no need of installing it. The interface will be implemented using the Treex::Web front-end (Sedlák, 2014).[11]

Acknowledgements

This research was supported by the grants FP7-ICT-2011-7-288487 (MosesCore), FP7-ICT-2013-10-610516 (QTLeap), GAUK 1572314, and SVV 260 104. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

Bibliography

Bojar, Ondřej. Analyzing Error Types in English-Czech Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:63–76, March 2011. ISSN 0032-6585.

Bojar, Ondrej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, 2013a.

[11] `https://ufal.mff.cuni.cz/tools/treex-web`

Bojar, Ondřej, Rudolf Rosa, and Aleš Tamchyna. Chimera – three heads for English-to-Czech translation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*, pages 92–98, Sofija, Bulgaria, 2013b. Bălgarska akademija na naukite, Association for Computational Linguistics.

Bojar, Ondrej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 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 2010. Association for Computational Linguistics. URL `http://www.statmt.org/wmt10/pdf/WMT03.pdf`.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan. 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 2011. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June 2012. Association for Computational Linguistics.

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

Fishel, Mark, Ondrej Bojar, and Maja Popovic. Terra: a collection of translation error-annotated corpora. In *LREC*, pages 7–14, 2012.

Hajič, Jan. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, 2004.

Hajič, Jan. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Hajičová, Eva, editor, *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pages 12–19. Prague Karolinum, Charles University Press, 1998.

Koehn, Philipp, 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. 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 2007. Association for Computational Linguistics.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: the Penn treebank. *Comp. Ling.*, 19:313–330, June 1993. ISSN 0891-2017.

Mareček, David, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. Two-step translation with grammatical post-processing. In Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 426–432, Edinburgh, UK, 2011. University of Edinburgh, Association for Computational Linguistics.

McDonald, Ryan, Koby Crammer, and Fernando Pereira. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 91–98. Association for Computational Linguistics, 2005.

Mohaghegh, Mahsa, Abdolhossein Sarrafzadeh, and Mehdi Mohammadi. A three-layer architecture for automatic post-editing system using rule-based paradigm. *WSSANLP-2013*, page 17, 2013.

Novák, Václav and Zdeněk Žabokrtský. Feature engineering in maximum spanning tree dependency parser. In Matoušek, Václav and Pavel Mautner, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 92–98, Pilsen, Czech Republic, 2007. Springer Science+Business Media Deutschland GmbH.

Och, Franz Josef and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

Popel, Martin and Zdeněk Žabokrtský. TectoMT: modular NLP framework. In *Proceedings of the 7th international conference on Advances in natural language processing*, IceTAL'10, pages 293–304, Berlin, Heidelberg, 2010. Springer-Verlag.

Rosa, Rudolf. Automatic post-editing of phrase-based machine translation outputs. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Praha, Czechia, 2013. URL `http://ufal.mff.cuni.cz/rudolf-rosa/master-thesis`.

Rosa, Rudolf. Depfix manual. Technical Report TR-2014-55, ÚFAL MFF UK, 2014a. URL `http://ufal.mff.cuni.cz/techrep/tr55.pdf`.

Rosa, Rudolf. MSTperl parser, 2014b. URL `http://hdl.handle.net/11858/00-097C-0000-0023-7AEB-4`.

Rosa, Rudolf and David Mareček. Dependency relations labeller for Czech. In Sojka, Petr, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue: 15th International Conference, TSD 2012. Proceedings*, number 7499 in Lecture Notes in Computer Science, pages 256–263, Berlin / Heidelberg, 2012. Masarykova univerzita v Brně, Springer Verlag.

Rosa, Rudolf, Ondřej Dušek, David Mareček, and Martin Popel. Using parallel features in parsing of machine-translated sentences for correction of grammatical errors. In *Proceedings of Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), ACL*, pages 39–48, Jeju, Korea, 2012a. Association for Computational Linguistics.

Rosa, Rudolf, David Mareček, and Ondřej Dušek. DEPFIX: A system for automatic correction of Czech MT outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368, Montréal, Canada, 2012b. Association for Computational Linguistics.

Rosa, Rudolf, David Mareček, and Aleš Tamchyna. Deepfix: Statistical post-editing of statistical machine translation using deep syntactic analysis. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 172–179,

Sofija, Bulgaria, 2013. Bălgarska akademija na naukite, Association for Computational Linguistics.

Sedlák, Michal. *Treex::Web*. Bachelor's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czechia, 2014. URL `https://lindat.mff.cuni.cz/services/treex-web/`.

Spoustová, Drahomíra, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha, 2007.

Straková, Jana, 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.

Stymne, Sara and Lars Ahrenberg. Using a grammar checker for evaluation and postprocessing of statistical machine translation. In *LREC*, 2010.

Tamchyna, Aleš, Martin Popel, Rudolf Rosa, and Ondřej Bojar. CUNI in WMT14: Chimera still awaits Bellerophon. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 195–200, Baltimore, MD, USA, 2014. Association for Computational Linguistics.

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

Zeman, Daniel, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. HamleDT: To parse or not to parse? In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

Address for correspondence:
Rudolf Rosa
`rosa@ufal.mff.cuni.cz`
Charles University in Prague,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
118 00 Praha 1, Czech Republic