Chimera – Three Heads for English-to-Czech Translation

Ondřej Bojar and **Rudolf Rosa** and **Aleš Tamchyna**
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
surname@ufal.mff.cuni.cz

Abstract

This paper describes our WMT submissions CU-BOJAR and CU-DEPFIX, the latter dubbed "CHIMERA" because it combines on three diverse approaches: TectoMT, a system with transfer at the deep syntactic level of representation, factored phrase-based translation using Moses, and finally automatic rule-based correction of frequent grammatical and meaning errors. We do not use any off-the-shelf system-combination method.

1 Introduction

Targeting Czech in statistical machine translation (SMT) is notoriously difficult due to the large number of possible word forms and complex agreement rules. Previous attempts to resolve these issues include specific probabilistic models (Subotin, 2011) or leaving the morphological generation to a separate processing step (Fraser et al., 2012; Mareček et al., 2011).

TectoMT (CU-TECTOMT, Galuščáková et al. (2013)) is a hybrid (rule-based and statistical) MT system that closely follows the analysis-transfer-synthesis pipeline. As such, it suffers from many issues but generating word forms in proper agreements with their neighbourhood as well as the translation of some diverging syntactic structures are handled well. Overall, TectoMT sometimes even ties with a highly tuned Moses configuration in manual evaluations, see Bojar et al. (2011).

Finally, Rosa et al. (2012) describes Depfix, a rule-based system for post-processing (S)MT output that corrects some morphological, syntactic and even semantic mistakes. Depfix was able to significantly improve Google output in WMT12, so now we applied it on an open-source system.

Our WMT13 system is thus a three-headed creature where, hopefully: (1) TectoMT provides missing word forms and safely handles some non-parallel syntactic constructions, (2) Moses exploits very large parallel and monolingual data, and boosts better lexical choice, (3) Depfix attempts to fix severe flaws in Moses output.

2 System Description

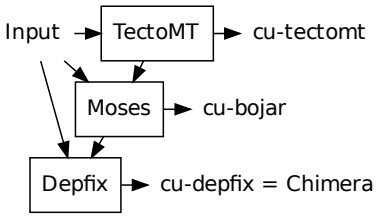

Figure 1: CHIMERA: three systems combined.

CHIMERA is a sequential combination of three diverse MT systems as depicted in Figure 1. Each of the intermediate stages of processing has been submitted as a separate primary system for the WMT manual evalution, allowing for a more thorough analysis.

Instead of an off-the-shelf system combination technique, we use TectoMT output as synthetic training data for Moses as described in Section 2.1 and finally we process its output using rule-based corrections of Depfix (Section 2.2). All steps directly use the source sentence.

2.1 Moses Setup for CU-BOJAR

We ran a couple of probes with reduced training data around the setup of Moses that proved successful in previous years (Bojar et al., 2012a).

2.1.1 Pre-processing

We use a stable pre-processing pipeline that includes normalization of quotation marks,[1] tokenization, tagging and lemmatization with tools

[1] We do not simply convert them to unpaired ASCII quotes but rather balance them and use other heuristics to convert most cases to the typographically correct form.

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 92–98,
Sofia, Bulgaria, August 8-9, 2013

Case	recaser	lc→form	utc	stc
BLEU	9.05	9.13	9.70	**9.81**

Table 1: Letter Casing

included in the Treex platform (Popel and Žabokrtský, 2010).

This year, we evaluated the end-to-end effect of truecasing. Ideally, English-Czech SMT should be trained on data where only names are uppercased (and neither the beginnings of sentences, nor all-caps headlines or exclamations etc). For these experiments, we trained a simple baseline system on 1 million sentence pairs from CzEng 1.0.

Table 1 summarizes the final (case-sensitive!) BLEU scores for four setups. The standard approach is to train SMT lowercase and apply a recaser, e.g. the Moses one, on the output. Another option (denoted "lc→form") is to lowercase only the source side of the parallel data. This more or less makes the translation model responsible for identifying names and the language model for identifying beginnings of sentences.

The final two approaches attempt at "truecasing" the data, i.e. the ideal lowercasing of everything except names. Our simple unsupervised truecaser ("utc") uses a model trained on monolingual data (1 million sentences in this case, same as the parallel training data used in this experiment) to identify the most frequent "casing shape" of each token type when it appears within a sentence and then converts its occurrences at the beginnings of sentences to this shape. Our supervised truecaser ("stc") casts the case of the *lemma* on the form, because our lemmatizers for English and Czech produce case-sensitive lemmas to indicate names. After the translation, only deterministic uppercasing of sentence beginnings is needed.

We confirm that "stc" as we have been using it for a couple of years is indeed the best option, despite its unpleasingly frequent omissions of names (incl. "Spojené státy", "the United States"). One of the rules in Depfix tries to cast the case from the source to the MT output but due to alignment errors, it is not perfect in fixing these mistakes.

Surprisingly, the standard recasing worked worse than "lc→form", suggesting that two Moses runs in a row are worse than one joint search.

We consider using a full-fledged named entity recognizer in the future.

		Tokens [M]	
Corpus	Sents [M]	English	Czech
CzEng 1.0	14.83	235.67	205.17
Europarl	0.65	17.61	15.00
Common Crawl	0.16	4.08	3.63

Table 2: Basic Statistics of Parallel Data.

2.1.2 Factored Translation for Morphological Coherence

We use a quite standard factored configuration of Moses. We translate from "stc" to two factors: "stc" and "tag" (full Czech positional morphological tag). Even though tags on the target side make the data somewhat sparser (a single Czech word form typically represents several cases, numbers or genders), we do not use any back-off or alternative decoding path. A high-order language model on tags is used to promote grammatically correct and coherent output. Our system is thus less prone to errors in local morphological agreement.

2.1.3 Large Parallel Data

The main source of our parallel data was CzEng 1.0 (Bojar et al., 2012b). We also used Europarl (Koehn, 2005) as made available by WMT13 organizers.[2] The English-Czech part of the new Common Crawl corpus was quite small and very noisy, so we did not include it in our training data. Table 2 provides basic statistics of the data.

Processing large parallel data can be challenging in terms of time and computational resources required. The main bottlenecks are word alignment and phrase extraction.

GIZA++ (Och and Ney, 2000) has been the standard tool for computing word alignment in phrase-based MT. A multi-threaded version exists (Gao and Vogel, 2008), which also supports incremental extensions of parallel data by applying a saved model on a new sentence pair. We evaluated these tools and measured their wall-clock time[3] as well as the final BLEU score of a full MT system.

Surprisingly, single-threaded GIZA++ was considerably faster than single-threaded MGIZA. Using 12 threads, MGIZA outperformed GIZA++ but the difference was smaller than we expected.

Table 3 summarizes the results. We checked the difference in BLEU using the procedure by Clark et al. (2011) and GIZA++ alignments were indeed

[2] http://www.statmt.org/wmt13/translation-task.html

[3] Time measurements are only indicative, they were affected by the current load in our cluster.

Alignment	Wallclock Time	BLEU
GIZA++	71	**15.5**
MGIZA 1 thread	114	15.4
MGIZA 12 threads	**51**	15.4

Table 3: Rough wallclock time [hours] of word alignment and the resulting BLEU scores.

Corpus	Sents [M]	Tokens [M]
CzEng 1.0	14.83	205.17
CWC Articles	36.72	626.86
CNC News	28.08	483.88
CNA	47.00	830.32
Newspapers	64.39	1040.80
News Crawl	24.91	444.84
Total	215.93	3631.87

Table 4: Basic Statistics of Monolingual Data.

little but significantly better than MGIZA in three MERT runs.

We thus use the standard GIZA++ aligner.

2.1.4 Large Language Models

We were able to collect a very large amount of monolingual data for Czech: almost 216 million sentences, 3.6 billion tokens. Table 4 lists the corpora we used. CWC Articles is a section of the Czech Web Corpus (Spoustová and Spousta, 2012). CNC News refers to a subset of the Czech National Corpus[4] from the news domain. CNA is a corpus of Czech News Agency stories from 1998 to 2012. Newspapers is a collection of articles from various Czech newspapers from years 1998 to 2002. Finally, News Crawl is the monolingual corpus made available by the organizers of WMT13.

We created an in-domain language model from all the corpora except for CzEng (where we only used the news section). We were able to train a 4-gram language model using KenLM (Heafield et al., 2013). Unfortunately, we did not manage to use a model of higher order. The model file (even in the binarized trie format with probability quantization) was so large that we ran out of memory in decoding.[5] We also tried pruning these larger models but we did not have enough RAM.

To cater for a longer-range coherence, we trained a 7-gram language model only on the News Crawl corpus (concatenation of all years). In this case, we used SRILM (Stolcke, 2002) and pruned n-grams so that (training set) model perplexity does not increase more than 10^{-14}. The data for this LM exactly match the domain of WMT test sets.

Token	Order	Sents [M]	Tokens [M]	ARPA.gz [GB]	Trie [GB]
stc	4	201.31	3430.92	28.2	11.8
stc	7	24.91	444.84	13.1	8.1
tag	10	14.83	205.17	7.2	3.0

Table 5: LMs used in CU-BOJAR.

Finally, we model sequences of morphological tags on the target side using a 10-gram LM estimated from CzEng. Individual sections of the corpus (news, fiction, subtitles, EU legislation, web pages, technical documentation and Navajo project) were interpolated to match WMT test sets from 2007 to 2011 best. This allows even out-of-domain data to contribute to modeling of overall sentence structure. We filtered the model using the same threshold 10^{-14}.

Table 5 summarizes the resulting LM files as used in CU-BOJAR and CHIMERA.

2.1.5 Bigger Tuning Sets

Koehn and Haddow (2012) report benefits from tuning on a larger set of sentences. We experimented with a down-scaled MT system to compare a couple of options for our tuning set: the default 3003 sentences of newstest2011, the default and three more Czech references that were created by translating from German, the default and two more references that were created by post-editing a variant of our last year's Moses system and also a larger single-reference set consisting of several newstest years. The preliminary results were highly inconclusive: negligibly higher BLEU scores obtained lower manual scores. Unable to pick the best configuration, we picked the largest. We tune our systems on "bigref", as specified in Table 6. The dataset consists of 11583 source sentences, 3003 of which have 4 reference translations and a subset (1997 sents.) of which has 2 reference translations constructed by post-editing. The dataset does not include 2010 data as a heldout for other foreseen experiments.

2.1.6 Synthetic Parallel Data

Galuščáková et al. (2013) describe several possibilities of combining TectoMT and phrase-based approaches. Our CU-BOJAR uses one of the simpler but effective ones: adding TectoMT output on the test set to our training data. As a contrast to

[4] http://korpus.cz/

[5] Due to our cluster configuration, we need to pre-load language models.

English	Czech	# Refs	# Snts
newstest2011	official + 3 more from German	4	3003
newstest2011	2 post-edits of a system similar to (Bojar et al., 2012a)	2	1997
newstest2009	official	1	2525
newstest2008	official	1	2051
newstest2007	official	1	2007
Total		4	11583

Table 6: Our big tuning set (bigref).

Test Set	Used in Training	Used in Tuning	Used in Final Eval
newstest07	*could have*	en+cs	–
newstest08	en+TectoMT	en+cs	–
newstest09	*could have*	en+cs	–
newstest10	en+TectoMT	*could have*	–
newstest11	en+TectoMT	en+cs	–
newstest12	en+TectoMT	en+cs	–
newstest13	en+TectoMT	–	en+cs

Table 7: Summary of test sets usage. "en" and "cs" denote the official English and Czech sides, resp. "TectoMT" denotes the synthetic Czech.

CU-BOJAR, we also examine PLAIN Moses setup which is identical but lacks the additional synthetic phrase table by TectoMT.

In order to select the best balance between phrases suggested by TectoMT and our parallel data, we provide these data as two separate phrase tables. Each phrase table brings in its own five-tuple of scores, one of which, the phrase-penalty functions as an indicator how many phrases come from which of the phrase tables. The standard MERT is then used to optimize the weights.[6,7]

We use one more trick compared to Galuščáková et al. (2013): we deliberately overlap our training and tuning datasets. When preparing the synthetic parallel data, we use the English side of newstests 08 and 10–13. The Czech side is always produced by TectoMT. We tune on bigref (see Table 6), so the years 08, 11 and 12 overlap. (We could have overlapped also years 07, 09 and 10 but we had them originally reserved for other purposes.) Table 7 summarizes the situation and highlights that our setup is fair: we never use the target side of our final evaluation set newstest2013. Some test sets are denoted "*could have*" as including them would still be correct.

The overlap allows MERT to estimate how useful are TectoMT phrases compared to the standard phrase table not just in general but on the specific foreseen test set. This deliberate overfitting to newstest 08, 11 and 12 then helps in translating newstest13.

This combination technique in its current state is rather expensive as a new phrase table is required for every new input document. However, if we fix the weights for the TectoMT phrase table, we can avoid re-tuning the system (whether this would degrade translation quality needs to be empirically evaluated). Moreover, if we use a dynamic phrase table, we could update it with TectoMT outputs on the fly, thus bypassing the need to retrain the translation model.

[6]Using K-best batch MIRA (Cherry and Foster, 2012) did not work any better in our setup.

[7]We are aware of the fact that Moses alternative decoding paths (Birch and Osborne, 2007) with similar phrase tables clutter n-best lists with identical items, making MERT less stable (Eisele et al., 2008; Bojar and Tamchyna, 2011). The issue was not severe in our case, CU-BOJAR needed 10 iterations compared to 3 iterations needed for PLAIN.

2.2 Depfix

Depfix is an automatic post-editing tool for correcting errors in English-to-Czech SMT. It is applied as a post-processing step to CU-BOJAR, resulting in the CHIMERA system. Depfix 2013 is an improvement of Depfix 2012 (Rosa et al., 2012).

Depfix focuses on three major types of language phenomena that can be captured by employing linguistic knowledge but are often hard for SMT systems to get right:

- morphological agreement, such as:
  - an adjective and the noun it modifies have to share the same morphological gender, number and case
  - the subject and the predicate have to agree in morphological gender, number and person, if applicable
- transfer of meaning in cases where the same meaning is expressed by different grammatical means in English and in Czech, such as:
  - a subject in English is marked by being a left modifier of the predicate, while in Czech a subject is marked by the nominative morphological case
  - English marks possessiveness by the preposition 'of', while Czech uses the genitive morphological case
  - negation can be marked in various ways in English and Czech
- verb-noun and noun-noun valency—see (Rosa et al., 2013)

Depfix first performs a complex lingustic anal-

System	BLEU	TER	WMT Ranking	
			Appraise	MTurk
CU-TECTOMT	14.7	0.741	0.455	0.491
CU-BOJAR	**20.1**	0.696	0.637	**0.555**
CU-DEPFIX	20.0	**0.693**	**0.664**	0.542
PLAIN Moses	19.5	0.713	–	–
GOOGLE TR.	–	–	0.618	0.526

Table 8: Overall results.

ysis of both the source English sentence and its translation to Czech by CU-BOJAR. The analysis includes tagging, word-alignment, and dependency parsing both to shallow-syntax ("analytical") and deep-syntax ("tectogrammatical") dependency trees. Detection and correction of errors is performed by rule-based components (the valency corrections use a simple statistical valency model). For example, if the adjective-noun agreement is found to be violated, it is corrected by projecting the morphological categories from the noun to the adjective, which is realized by changing their values in the Czech morphological tag and generating the appropriate word form from the lemma-tag pair using the rule-based generator of Hajič (2004).

Rosa (2013) provides details of the current version of Depfix. The main additions since 2012 are valency corrections and lost negation recovery.

3 Overall Results

Table 8 reports the scores on the WMT13 test set. BLEU and TER are taken from the evaluation web site[8] for the *normalized* outputs, case insensitive. The normalization affects typesetting of punctuation only and greatly increases automatic scores. "WMT ranking" lists results from judgments from Appraise and Mechanical Turk. Except CU-TECTOMT, the manual evaluation used non-normalized MT outputs. The figure is the WMT12 standard interpretation as suggested by Bojar et al. (2011) and says how often the given system was ranked better than its competitor across all 18.6k non-tying pairwise comparisons extracted from the annotations.

We see a giant leap from CU-TECTOMT to CU-BOJAR, confirming the utility of large data. However, CU-TECTOMT had something to offer since it improved over PLAIN, a very competitive baseline, by 0.6 BLEU absolute. Depfix seems to slightly worsen BLEU score but slightly improve TER; the manual evaluation is similarly indecisive.

[8] http://matrix.statmt.org/

System	# Tokens	% Tokens
All	22920	76.44
Moses	3864	12.89
TectoMT	2323	7.75
Other	877	2.92

Table 9: CHIMERA components that contribute "confirmed" tokens.

System	# Tokens	% Tokens
None	21633	79.93
Moses	2093	7.73
TectoMT	2585	9.55
Both	385	1.42
CU-BOJAR	370	1.37

Table 10: Tokens missing in CHIMERA output.

4 Combination Analysis

We now closely analyze the contributions of the individual engines to the performance of CHIMERA. We look at translations of the newstest2013 sets produced by the individual systems (PLAIN, CU-TECTOMT, CU-BOJAR, CHIMERA).

We divide the newstest2013 reference tokens into two classes: those successfully produced by CHIMERA (Table 9) and those missed (Table 10). The analysis can suffer from false positives as well as false negatives, a "confirmed" token can violate some grammatical constraints in MT output and an "unconfirmed" token can be a very good translation. If we had access to more references, the issue of false negatives would decrease.

Table 9 indicates that more than 3/4 of tokens confirmed by the reference were available in all CHIMERA components: PLAIN Moses, CU-TECTOMT alone but also in the subsequent combinations CU-BOJAR and the final CU-DEPFIX.

PLAIN Moses produced 13% tokens that TectoMT did not provide and TectoMT output roughly 8% tokens unknown to Moses. However, note that it is difficult to distinguish the effect of different model weights: PLAIN *might have* produced some of those tokens as well if its weights were different. The row "Other" includes cases where e.g. Depfix introduced a confirmed token that none of the previous systems had.

Table 10 analyses the potential of CHIMERA components. These tokens from the reference were *not* produced by CHIMERA. In almost 80% of cases, the token was not available in any 1-best output; it *may* have been available in Moses phrase

tables or the input sentence.

TectoMT offered almost 10% of missed tokens, but these were not selected in the subsequent combination. The potential of Moses is somewhat lower (about 8%) because our phrase-based combination is likely to select wordings that score well in a phrase-based model. 385 tokens were suggested by both TectoMT and Moses alone, but the combination in CU-BOJAR did not select them, and finally 370 tokens were produced by the combination while they were *not* present in 1-best output of neither TectoMT nor Moses. Remember, all these tokens eventually did not get to CHIMERA output, so Depfix must have changed them.

4.1 Depfix analysis

Table 11 analyzes the performance of the individual components of Depfix. Each *evaluated* sentence was either *modified* by a Depfix component, or not. If it was *modified*, its quality could have been evaluated as better (*improved*), worse (*worsened*), or the same (*equal*) as before. Thus, we can evaluate the performance of the individual components by the following measures:[9]

$$precision = \frac{\#improved}{\#improved+\#worsened} \quad (1)$$

$$impact = \frac{\#modified}{\#evaluated} \quad (2)$$

$$useless = \frac{\#equal}{\#modified} \quad (3)$$

Please note that we make an assumption that if a sentence was modified by multiple Depfix components, they all have the same effect on its quality. While this is clearly incorrect, it is impossible to accurately determine the effect of each individual component with the evaluation data at hand. This probably skews especially the reported performance of "high-impact" components, which often operate in combination with other components.

The evaluation is computed on 871 hits in which CU-BOJAR and CHIMERA were compared.

The results show that the two newest components – Lost negation recovery and Valency model – both modify a large number of sentences. Valency model seems to have a slightly *negative* effect on the translation quality. As this is the only statistical component of Depfix, we believe that this is caused by the fact that its parameters were not tuned on the final CU-BOJAR system, as the tuning has to be done semi-manually and the final system was not available in advance. On the other hand, Lost negation recovery seems to have a highly positive effect on translation quality. This is to be expected, as a lost negation often leads to the translation bearing an opposite meaning to the original one, which is probably one of the most serious errors that an MT system can make.

[9]We use the term *precision* for our primary measure for convenience, even though the way we define it does not match exactly its usual definition.

Depfix component	Prc.	Imp.	Usl.
Aux 'be' agr.	–	1.4%	100%
No prep. without children	–	0.5%	100%
Sentence-initial capitalization	0%	0.1%	0%
Prepositional morph. case	0%	2.1%	83%
Preposition - noun agr.	40%	3.8%	70%
Noun number projection	41%	7.2%	65%
Valency model	48%	10.6%	66%
Subject - nominal pred. agr.	50%	3.8%	76%
Noun - adjective agr.	55%	17.8%	75%
Subject morph. case	56%	8.5%	57%
Tokenization projection	56%	3.0%	38%
Verb tense projection	58%	5.2%	47%
Passive actor with 'by'	60%	1.0%	44%
Possessive nouns	67%	0.9%	25%
Source-aware truecasing	67%	2.8%	50%
Subject - predicate agr.	68%	5.1%	57%
Pro-drop in subject	73%	3.4%	63%
Subject - past participle agr.	75%	6.3%	42%
Passive - aux 'be' agr.	77%	4.8%	69%
Possessive with 'of'	78%	1.5%	31%
Present continuous	78%	1.5%	31%
Missing reflexive verbs	80%	1.6%	64%
Subject categories projection	83%	3.7%	62%
Rehang children of aux verbs	83%	5.5%	62%
Lost negation recovery	90%	7.2%	38%

Table 11: Depfix components performance analysis on 871 sentences from WMT13 test set.

5 Conclusion

We have reached our chimera to beat Google Translate. We combined all we have: a deep-syntactic transfer-based system TectoMT, very large parallel and monolingual data, factored setup to ensure morphological coherence, and finally Depfix, a rule-based automatic post-editing system that corrects grammaticality (agreement and valency) of the output as well as some features vital for adequacy, namely lost negation.

Acknowledgments

This work was partially supported by the grants P406/11/1499 of the Grant Agency of the Czech Republic, FP7-ICT-2011-7-288487 (MosesCore) and FP7-ICT-2010-6-257528 (Khresmoi) of the European Union and by SVV project number 267 314.

References

Alexandra Birch and Miles Osborne. 2007. CCG Supertags in Factored Statistical Machine Translation. In *In ACL Workshop on Statistical Machine Translation*, pages 9–16.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In *Proc. of WMT*, pages 330–336. ACL.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proc. of WMT*, pages 1–11. ACL.

Ondřej Bojar, Bushra Jawaid, and Amir Kamran. 2012a. Probes in a Taxonomy of Factored Phrase-Based Models. In *Proc. of WMT*, pages 253–260. ACL.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012b. The Joy of Parallelism with CzEng 1.0. In *Proc. of LREC*, pages 3921–3928. ELRA.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proc. of NAACL/HLT*, pages 427–436. ACL.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. of ACL/HLT*, pages 176–181. ACL.

Andreas Eisele, Christian Federmann, Hervé Saint-Amand, Michael Jellinghaus, Teresa Herrmann, and Yu Chen. 2008. Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System. In *Proc. of WMT*, pages 179–182. ACL.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling Inflection and Word-Formation in SMT. In *Proc. of EACL 2012*. ACL.

Petra Galuščáková, Martin Popel, and Ondřej Bojar. 2013. PhraseFix: Statistical Post-Editing of TectoMT. In *Proc. of WMT13*. Under review.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57. ACL.

Jan Hajič. 2004. *Disambiguation of rich inflection: computational morphology of Czech*. Karolinum.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proc. of ACL*.

Philipp Koehn and Barry Haddow. 2012. Towards Effective Use of Training Data in Statistical Machine Translation. In *Proc. of WMT*, pages 317–321. ACL.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step translation with grammatical post-processing. In *Proc. of WMT*, pages 426–432. ACL.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *ACL*. ACL.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *IceTAL 2010*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304. Iceland Centre for Language Technology (ICLT), Springer.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A system for automatic correction of Czech MT outputs. In *Proc. of WMT*, pages 362–368. ACL.

Rudolf Rosa, David Mareček, and Aleš Tamchyna. 2013. Deepfix: Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis. Bălgarska akademija na naukite, ACL.

Rudolf Rosa. 2013. Automatic post-editing of phrase-based machine translation outputs. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Praha, Czechia.

Johanka Spoustová and Miroslav Spousta. 2012. A High-Quality Web Corpus of Czech. In *Proc. of LREC*. ELRA.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904.

Michael Subotin. 2011. An exponential translation model for target language morphology. In *Proc. of ACL/HLT*, pages 230–238. ACL.