2010 Failures in English-Czech Phrase-Based MT *

Ondřej Bojar and Kamil Kos
Charles University in Prague, Institute of Formal and Applied Linguistics (ÚFAL)
Malostranské náměstí 25, Praha 1, CZ-11800, Czech Republic
bojar@ufal.mff.cuni.cz, kamilkos@email.cz

Abstract

The paper describes our experiments with English-Czech machine translation for WMT10[1] in 2010. Focusing primarily on the translation to Czech, our additions to the standard Moses phrase-based MT pipeline include two-step translation to overcome target-side data sparseness and optimization towards SemPOS, a metric better suited for evaluating Czech. Unfortunately, none of the approaches bring a significant improvement over our standard setup.

1 Introduction

Czech is a flective language with very rich morphological system. Translation between Czech and English poses different challenges for each of the directions.

When translating from Czech, the word order usually needs only minor changes (despite the issue of non-projectivity, a phenomenon occurring at 2% of words but in 23% of Czech sentences, see Hajičová et al. (2004) and Holan (2003)). A much more severe issue is caused by the Czech vocabulary size. Fortunately, this can be to a certain extent mitigated by backing-off to Czech lemmas if the exact forms are not available.

We are primarily interested in the harder task of translating to Czech and most of the paper deals with this direction. After a brief specification of data sets, pre-processing and evaluation method in this section, we provide details on the issue of Czech vocabulary size (Section 2). We describe our current attempts at generating Czech word forms in Section 3. Partly due to the large vocabulary size of Czech, BLEU score (Papineni et al., 2002) correlates rather poorly with human judgments. We summarize our efforts to use a better metric in the model optimization in Section 4. The final Section 5 lists the exact configurations of our English↔Czech primary submissions for WMT10, including the back-off to lemmas we use for Czech-to-English.

1.1 Data and Pre-Processing Pipeline

Throughout the paper, we use CzEng 0.9 (Bojar and Žabokrtský, 2009)[2] as our main parallel corpus. Following CzEng authors' request, we did not use sections 8* and 9* reserved for evaluation purposes.

As the baseline training dataset ("Small" in the following) only the news domain of CzEng (126k parallel sentences) is used. For large-scale experiments ("Large" in the following) and our primary WMT10 submissions, we use all CzEng domains except `navajo` and add the EMEA corpus (Tiedemann, 2009)[3,4] of 7.5M parallel sententes.

As our monolingual data we use by default only the target side of the parallel corpus. For experiments reported here, we also use the monolingual data provided by WMT10 organizers for Czech. Our primary WMT10 submission includes further monolingual data, see Section 5.1.

We use a slightly modified tokenization rules compared to CzEng export format. Most notably, we normalize English abbreviated negation and auxiliary verbs ("couldn't" → "could not") and attempt at normalizing quotation marks to distinguish between the opening and closing one follow-

*The work on this project was supported by the grants EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), GAČR P406/10/P259, and MSM 0021620838. Thanks to David Kolovratník for the help with manual evaluation.

[1] `http://www.statmt.org/wmt10/`

[2] `http://ufal.mff.cuni.cz/czeng`

[3] `http://urd.let.rug.nl/tiedeman/OPUS`

[4] Unfortunately, the EMEA corpus is badly tokenized on the Czech side. Most frequently, fractional numbers are split into several tokens (e.g. "3, 14"). We attempted to reconstruct the original detokenized form using a small set of regular expressions.

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 60–66,
Uppsala, Sweden, 11-16 July 2010.

	Large	Small	Dev
Sents	7.5M	126.1k	2.5k
Czech Tokens	79.2M	2.6M	55.8k
English Tokens	89.1M	2.9M	49.9k
Czech Vocabulary	923.1k	138.7k	15.4k
English Vocabulary	646.3k	64.7k	9.4k
Czech Lemmas	553.5k	60.3k	9.5k
English Lemmas	611.4k	53.8k	7.7k

Table 1: Corpus and vocabulary sizes.

ing proper typesetting rules.

The rest of our pre-processing pipeline matches the processing employed in CzEng (Bojar and Žabokrtský, 2009).[5] We use "supervised truecasing", meaning that we cast the case of the lemma to the form, relying on our morphological analyzers and taggers to identify proper names, all other words are lowercased.

The differences in relations between Czech and English Large and Small datasets can be attributed either to domain differences or possibly due to noise in CzEng.

1.2 Evaluation

We use WMT10 development sets for tuning (news-test2008) and evaluation (news-test2009). The official scores on news-test2010 are given only in the main WMT10 paper and not here.

The BLEU scores reported in this paper are based on truecased word forms in the original tokenization as provided by the decoder. Therefore they are likely to differ from figures reported elsewhere.

The $\pm$ value given with each BLEU score is the average of the distances to the lower and upper empirical 95% confidence bounds estimated using bootstrapping (Koehn, 2004).

2 Issues of Czech Vocabulary Size

Table 1 summarizes the differences of Czech and English vocabulary sizes in our parallel corpora. We see that the vocabulary size of Czech forms (truecased) is more than double compared to English in the Small dataset and significantly larger in the Large dataset as well. On the other hand, the number of distinct Czech and English lemmas is nearly identical.

[5]Due to the subsequent processing, incl. parsing, the tokenization of English follows PennTreebenk style. The rather unfortunate convention of treating hyphenated words as single tokens increases our out-of-vocabulary rate. Next time, we will surely post-tokenize the parsed text.

	Distortion Limit				
TOpts	3	6	10	30	40
1	0.2	0.3	0.3	0.3	0.3
5	0.8	0.9	1.0	1.0	1.0
10	1.1	1.3	1.5	1.5	1.5
20	1.2	1.5	1.7	1.7	1.7
50	1.2	1.5	1.7	1.7	1.7
100	1.2	1.5	1.7	1.7	1.7

Table 3: Percentage of sentences reachable in Czech-to-English small setting with various distortion limits and translation options per coverage (TOpts) (BLEU score 14.76$\pm$0.44).

2.1 Out-of-Vocabulary Rates

Table 2 lists out-of-vocabulary (OOV) rates of our Small and Large data setting given the development corpus. We calculate the rates for both the complete corpus and the restricted set of phrases extracted from the corpus. (Note that higher-order n-gram rates are estimated using phrases as independent units, no combination of phrases is performed.) We also list the effective OOV rate for English-to-Czech translation where all (English) words from each source sentence can be also produced in the hypothesis.

We see that in the small setting, the OOV rate is almost double for Czech than for English. The OOV is significantly decreased by enlarging the corpus or lemmatizing the word forms.

If we consider only the words available in the phrase tables, the issue of Czech with limited data is striking: 10–12% of devset tokens are not available in the training data.

2.2 Reachability of Training and Reference Translations

Schwartz (2008) extended Moses to support "constraint decoding", that is to perform an exhaustive search through the space of hypotheses in order to reach the reference translation (and get its score).

The current implementation of the exhaustive search in Moses is in fact subject to several configuration parameters, most importantly the number of translation options considered for each span (`-max-trans-opt-per-coverage`) and the distortion limit (`-distortion-limit`).

Given his aim, Schwartz (2008) uses the output of four MT systems translating from different languages to English as the references and notes that only around 10% of the reference translations are reachable by an independent Swedish-English MT system.

		n-grams Out of Corpus Voc.				n-grams Out of Phrase-Table Voc.			
Dataset	Language	1	2	3	4	1	2	3	4
Large	Czech	2.2%	30.5%	70.2%	90.3%	3.9%	44.1%	82.2%	95.6%
Large	English	1.5%	13.7%	47.3%	78.8%	2.1%	22.4%	63.5%	89.1%
Large	Czech + English input sent	1.5%	29.4%	69.6%	90.1%	3.1%	42.8%	81.5%	95.3%
Small	Czech	6.7%	48.1%	83.0%	95.5%	12.5%	65.4%	91.9%	98.6%
Small	English	3.6%	28.1%	68.3%	90.9%	6.3%	45.4%	84.3%	97.0%
Small	Czech + English input sent	5.2%	46.6%	82.4%	95.2%	10.6%	63.7%	91.2%	98.3%
Small	Czech lemmas	4.1%	36.3%	75.8%	92.8%	5.8%	52.6%	87.7%	97.4%
Small	English lemmas	3.4%	24.6%	64.6%	89.4%	6.9%	53.2%	87.9%	97.5%
Small	Czech + English input sent lemmas	3.1%	35.7%	75.6%	92.8%	5.1%	38.1%	80.8%	96.2%

Table 2: Out-of-vocabulary rates.

	Distortion Limit				
TOpts	3	6	10	30	40
1	0.4	0.4	0.4	0.4	0.4
5	1.5	1.9	2.0	2.0	2.0
10	2.5	3.2	3.5	3.5	3.5
20	3.7	5.0	5.5	5.6	5.6
50	4.9	6.7	8.0	8.6	8.6
100	5.3	7.6	9.1	9.4	9.4

Table 4: Percentage of sentences reachable in Czech-to-English large setting, two alternative decoding paths to translate from Czech lemma if the form is not available in the translation table (BLEU score 18.70±0.46).

We observe that reaching man-made reference translations in Czech-to-English translation is far harder. Table 3 provides the figures for small data setting (and no phrase table filtering). The best reachability we can hope for is given in Table 4 where we allow to use source word lemmas if the exact form is not available. We see that the default limits (50 translation options per span and distortion limit of 6) leave us with only 6.7% sentences reachable.

While not directly important for your training, the figures still underpin the issue of sparse data in Czech-English translation.

3 Targetting Czech Word Forms

Bojar (2007) experimented with several translation scenarios, including what we will call MorphG, i.e. the independent translation of lemma to lemma and tag to tag followed by a generation step to produce target-side word form. With the small training set available then, the MorphG model performed equally well as a simpler direct translation followed by target-side tagging and an additional n-gram model over morphological tags. Koehn and Hoang (2007) reports even a large loss with MorphG for German-to-English if the alternative of direct form-to-form translation is not available.

Bojar et al. (2009b) applied the two alternative decoding paths (direct form-to-form and MorphG, labelled "T+C+C&T+T+G") to English-Czech but they were able to use only 84k sentences. For the full training set of 2.2M sentences, the model was too big to fit in reasonable disk limits. More importantly, already in the small data setting, the complex model suffered from little stability due to abundance of features (5 features per phrase-table plus tree features for three LMs), so nearly the same performance on the development set gave largely varying quality on the independent test set.

The most important issue of the MorphG setup, however, is the explosion of translation options. Due to the "synchronous factors" approach of Moses (Koehn and Hoang, 2007), all translation options have to be fully constructed before the main search begins. The MorphG model however licenses too many possible combinations of lemmas, tags and final word forms, so the pruning of translation options strikes hard, causing search errors. For more details, see Bojar et al. (2009a) where a similar issue occurs for treelet-based translation.

3.1 Two-Step Translation

In order to avoid the explosion of the translation options[6], we experimented with two-step translation.

The first step translates from English to lemmatized Czech augmented to preserve important semantic properties known from the source phrase. The second step is a monotone translation from the lemmas to fully inflected Czech. The idea behind the delimitation is that all the morphological properties of Czech words that can be established

[6] and also motivated when we noticed that reading MT output to *lemmatized* Czech is sometimes more pleasant and informative than regular phrase-based output

Data Size		Simple		Two-Step	
Parallel	Mono	BLEU	SemPOS	BLEU	SemPOS
Small	Small	10.28±0.40	29.92	10.38±0.38	30.01
Small	Large	12.50±0.44	31.01	12.29±0.47	31.40
Large	Large	14.17±0.51	33.07	14.06±0.49	32.57

Table 5: Performance of direct (Simple) and two-step factored translation in small and large data setting.

regardless the English source should not cause parallel data sparseness and clutter the search. Instead, they should be decided based on context in the second phase only.

Specifically, the intermediate Czech represents most words as tuples containing only: lemma, negation, grade (of adjectives and adverbs), number (of nouns, adjectives, verbs) and detailed part of speech (constraining also e.g. verb tense of Czech verbs). Some words are handled separately:

- Pronouns, punctuation and the verbs "být" (to be) and "mít" (to have) are represented using their lowecased full forms because they are very frequent, often auxiliary to other words and their exact form best captures the available and necessary detail of many morphological and syntactic properties.
- Prepositions are represented using their lemmas and case because the case of a noun phrase is actually introduced by the governing word (e.g. the verb that subcategorized for the noun phrase or the preposition for prepositional phrases).

Table 5 compares the scores of the simple phrase-based and the two-step translation via augmented Czech lemmas as described above. The small and large parallel data denote the datasets described in Section 1.1. The small monolingual set means just the news domain of CzEng, while the large monolingual set means WMT10 monolingual Czech texts (and no CzEng data). Note that the monolingual data serve three purposes in the two-step approach: the language model for the first phase, the translation model in the second phase (monotone and restricted to phrase-length of 1; longer phrases did not bring significant improvement either), and the language model of the second phase. Ignoring the opportunity to use the monolingual set as the language model in the first phase already hurts the performance.

We see that the results as evaluated both by BLEU and SemPOS (see Section 4 below) are rather mixed but not that surprising. There is a negligible gain in the Small-Small setting, a mixed outcome in the Small-Large and a little loss in the Large-Large setting.

	Two- -Step	Both Fine	Both Wrong	Simple	Total
Two-Step	**23**	4	8	-	**35**
Both Fine	7	14	17	5	43
Both Wrong	8	1	28	2	39
Simple	-	3	7	**23**	33
Total	**38**	22	60	30	150

Table 6: Manual micro-evaluation of Simple (12.50±0.44) vs. Two-step (12.29±0.47) model in the Small-Large setting.

The most interesting result is the Small-Large setting: BLEU (insignificantly) prefers the simple and SemPOS the two-step model. It thus seems that a large target-side LM is sufficient to improve the BLEU score, despite the untackled issue of bilingual data sparseness.

We carried out a quick manual evaluation of 150 sentences by two annotators (one of the authors and a third person; systems anonymized): for each input segment, either one of the outputs is distinguishably better or both are equally wrong or equally acceptable. As listed in the confusion matrix in Table 6, each annotator independently marginally prefers the two-step approach but the intersection does not confirm that.[7] One good thingis that the annotators do not completely contradict each other's preference.

Ultimately, we did not use the two-step approach in our primary submission, but we feel there is still some unexploited potential in this phrase-based approximation of the technique separating properties of words handled in the translation phase from properties implied by the target-side (grammatical) context only. Certainly, the representation of the intermediate language can

[7] Of the 23 sentences improved by the two-step setup, about three quarters indeed had an improvement in lexical coverage or better morphological choice of a word. Of the 23 sentences where the two-step model hurts, about a half suffered from errors related to superfluous auxiliary words in Czech that seem to be introduced by a bias towards word-for-word translation. This bias is not inherent to the model, only the (normalized) phrase penalty weight happened to get nearly three times bigger than in the simple model.

be still improved, and more importantly, the second phase of monotone decoding could be handled by a more appropriate model capable of including more additional (source) context features.[8]

4 Optimizing towards SemPOS

In our setup, we use minimum error-rate training (MERT, Och (2003)) to optimize weights of model components. In the standard implementation in Moses, BLEU (Papineni et al., 2002) is used as the objective function, despite its rather disputable correlation with human judgments of MT quality.

Kos and Bojar (2009) introduced SemPOS, a metric that performs much better in terms of correlation to human judgments when translating to Czech. Naturally, we wanted to optimize towards SemPOS.

SemPOS computes the overlapping of autosemantic (content-bearing) word lemmas in the candidate and reference translations given a fine-grained semantic part of speech (sempos[9]), as defined in Hajič et al. (2006), and outputs average overlapping score over all sempos types.

The SemPOS metric outperformed common metrics as BLEU, TER (Snover et al., 2006) or an adaptation of Meteor (Lavie and Agarwal, 2007) for Czech on test sets from WMT08 (Callison-Burch et al., 2008).

4.1 Integrating SemPOS to MERT

In our experiments we used Z-MERT (Zaidan, 2009), a recent implementation of the MERT algorithm, to optimize model parameters.

The SemPOS metric requires to remove all auxiliary words and to identify the (deep-syntactic) lemmas and semantic part of speech for autosemantic words. When employed in MERT training, the whole n-best list of candidates has to processed like this at each iteration.

We use the TectoMT platform (Žabokrtský and Bojar, 2008)[10] for the linguistic processing. TectoMT follows the complete pipeline of tagging, surface-syntactic analysis and deep-syntactic analysis, which is the best but rather costly way to obtain the required information.

Therefore, we use two different ways of obtaining lemmas and semantic parts of speech in the MERT loop:

- indeed apply TectoMT processing to the n-best list at each iteration (parallelized to 15 CPUs),
- apply TectoMT to the *training data*, express the (deep) lemma and sempos as additional factors using a blank value for auxiliary words, and using Moses factored translation to translate from English forms to triplets of Czech form, deep lemma and sempos.

	BLEU	SemPOS	Iters	Time
TectoMT	10.11±0.40	29.69	20	2d12.0h
in MERT	9.53±0.39	29.69	10	1d12.0h
Factored	9.46±0.37	29.36	10	2.4h
translation	8.20±0.37	29.68	-	-
	6.96±0.33	27.79	9	1.7h

Table 7: Five independent MERT runs optimizing towards SemPOS with semantic parts of speech and lemmas provided either by TectoMT on the fly or by Moses factored translation.

Table 7 lists several ZMERT runs when optimizing a simple form→form phrase-based model (small data setting) towards SemPOS. One observation is that using TectoMT in the MERT loop is unbearably costly and we avoided it in the subsequent experiments. More importantly, from the huge differences in the final BLEU as well as SemPOS scores (evaluated on the independent test set), we see how unstable the search is.

SemPOS, while good at comparing different MT systems, is very bad at comparing candidates from a single system in an n-best list. This can be easily explained by its low sensitivity to precision: SemPOS disregards word forms as well as all auxiliary words. This is a good thing to compare very different candidates (where each of the systems already struggled to produce a coherent output) but is of very little help when comparing candidates of a single system, because these candidates tend to differ rather in forms than in lexical choice.

4.2 Combination of SemPOS and BLEU

To compensate for some of the shortcomings of SemPOS, we also attempted to optimize towards a linear combination of SemPOS and BLEU. This should increase the suitability of the metric for MERT optimization because BLEU will take correct word forms into account while SemPOS should promote better lexical choice (possibly not confirmed by BLEU due to a different word form than in the reference).

Table 8 provides the results of various weight

[8] We are grateful to Trevor Cohn for the suggestion.

[9] In the following text we will use SemPOS to denote the SemPOS metric. When speaking about the semantic part of speech, we will write sempos type or sempos tag.

[10] `http://ufal.mff.cuni.cz/tectomt/`

W.	BLEU	SemPOS	W.	BLEU	SemPOS
1:0	10.42±0.38	29.91	3:1	10.30±0.39	30.03
1:1	10.15±0.39	29.81	10:1	10.17±0.40	29.58
1:1	9.42±0.37	29.30	1:2	10.11±0.38	29.80
2:1	10.37±0.38	29.95	1:10	9.44±0.40	29.74

Table 8: Optimizing towards a linear combination of BLEU and SemPOS (weights in this order), small data setting.

	BLEU	SemPOS
BLEU alone	14.08±0.50	32.44
SemPOS-BLEU (1:1)	13.79±0.55	33.17

Table 9: Optimizing towards BLEU and/or SemPOS in large data setting.

settings, including the optimization towards BLEU alone using ZMERT implementation. We see that the stability is much better, only few runs suffered a minor loss (including 1:1 in one case). Unfortunately, the differences in final BLEU and SemPOS scores are all within confidence intervals when trained on the small dataset.

Table 9 documents that in our large data setting, MERT indeed achieves slightly higher SemPOS (and lower BLEU) when optimizing towards it. This corresponds with the intuition that with more variance in lexical choices available in the phrase tables, SemPOS can help to balance model features. The current set of weights is rather limited, so our future experiments should focus on actually providing means to e.g. domain adaptation by using features indicating the applicability of a phrase in a specific domain.

5 Our Primary Submissions to WMT10

5.1 English-to-Czech Translation

Given the little or no improvements achieved by the many configurations we tried, our English-to-Czech primary submission is rather simple:

- Standard GIZA++ word alignment based on both source and target lemmas.
- Two alternative decoding paths; forms always truecased: form+tag→form & form→form.
  The first path is more specific and helps to preserve core syntactic elements in the sentence. Without the tag, ambiguous English words could often all translate as e.g. nouns, leading to no verb in the Czech sentence. The default path serves as a back-off.
- Significance filtering of the phrase tables (Johnson et al., 2007) implemented for Moses by Chris Dyer; default settings of filter value `a+e` and the cut-off 30.
- Two separate 5-gram Czech LMs of truecased forms each of which interpolates models trained on the following datasets; the interpolation weights were set automatically using SRILM (Stolcke, 2002) based on the target side of the development set:[11]
  - Interpolated CzEng domains: news, web, fiction. The rationale behind the selection of the domains is that we prefer prose-like texts for LM estimation (and not e.g. technical documentation) while we want as much parallel data as possible.
  - Interpolated monolingual corpora: WMT09 monolingual, WMT10 monolingual, Czech National Corpus (Kocek et al., 2000) sections SYN2000+2005+2006PUB.
- Lexicalized reordering (`or-bi-fe`) based on forms.
- Standard Moses MERT towards BLEU.

	Large	Small
Backed-off by source lemmas	18.95±0.45	14.95±0.48
form→form only	18.41±0.44	14.73±0.47

Table 10: Translation from Czech better when backed-off by source lemmas.

5.2 Czech-to-English Translation

For Czech-to-English translation we experimented with far fewer configuration options. Our primary submission is configured as follows:

- Two alternative decoding paths; forms always truecased: form→form & lemma→form.
- Significance filtering as in Section 5.1.
- 5-gram English LM based on CzEng English side only.[12]
- Lexicalized reordering (`or-bi-fe`) based on forms.
- Standard Moses MERT towards BLEU.

Table 10 documents the utility of the additional decoding path from Czech lemmas in both small and large setting, surprisingly less significant in the small setting. Later experiments with system combination by Kenneth Heafield indicated that while our system is not among the top three, it brings an advantage to the combination.

6 Conclusion

We provided an extensive documentation of Czech data sparseness issue for machine translation. We attempted to tackle the problem of constructing the target-side form by a two-step translation setup and the problem of unreliable automatic evaluation by employing a new metric in MERT loop, neither with much success so far. Both of the attempts however deserve further exploration. Additionally, we provide the exact configurations of our WMT10 primary submissions.

[11] The subsequent MERT training using the same development test may suffer from overestimating the language model weights, but we did not observe the issue, possibly due to only moderate overlap of the datasets.

[12] We attempted to use a second LM trained on English Gigaword by Chris Callison-Burch, but we observed a drop in BLEU score from 18.95±0.45 to 18.03±0.44 probably due to different tokenization guidelines applied.

References

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

Ondřej Bojar, Miroslav Janíček, and Miroslav Týnovský. 2009a. Evaluation of Tree Transfer System. Project Euromatrix - Deliverable 3.4, Institute of Formal and Applied Linguistics, Charles University in Prague.

Ondřej Bojar, David Mareček, Václav Novák, Martin Popel, Jan Ptáček, Jan Rouš, and Zdeněk Žabokrtský. 2009b. English-Czech MT in 2008. In *Proc. of Fourth Workshop on Statistical Machine Translation, ACL*, Athens, Greece.

Ondřej Bojar. 2007. English-to-Czech Factored Machine Translation. In *Proc. of the Second Workshop on Statistical Machine Translation, ACL*, Prague, Czech Republic, June.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proc. of the Third Workshop on Statistical Machine Translation, ACL*, Columbus, Ohio.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. 2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, and Daniel Zeman. 2004. Issues of Projectivity in the Prague Dependency Treebank. *The Prague Bulletin of Mathematical Linguistics*, 81.

Tomáš Holan. 2003. K syntaktické analýze českých(!) vět. In *MIS 2003*. MATFYZPRESS.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proc. of EMNLP-CoNLL*, Prague, Czech Republic.

Jan Kocek, Marie Kopřivová, and Karel Kučera, editors. 2000. *Český národní korpus - úvod a příručka uživatele*. FF UK - ÚČNK, Prague.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proc. of EMNLP*.

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

Kamil Kos and Ondřej Bojar. 2009. Evaluation of Machine Translation Metrics for Czech as the Target Language. *The Prague Bulletin of Mathematical Linguistics*, 92.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proc. of the Second Workshop on Statistical Machine Translation, ACL*, Prague, Czech Republic.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of ACL*, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*, Philadelphia, Pennsylvania.

Lane Schwartz. 2008. Multi-source translation methods. In *Proc. of AMTA*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of AMTA*.

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

Jörg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Proc. of Recent Advances in NLP (RANLP)*.

Zdeněk Žabokrtský and Ondřej Bojar. 2008. TectoMT, Developer's Guide. Technical Report TR-2008-39, Institute of Formal and Applied Linguistics, Charles University in Prague.

Omar F. Zaidan. 2009. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. *The Prague Bulletin of Mathematical Linguistics*, 91.