View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Biblio at Institute of Formal and Applied Linguistics

# Improving Evaluation of English-Czech MT through Paraphrasing

**Petra Barančíková, Rudolf Rosa, Aleš Tamchyna**

Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
{barancikova,rosa,tamchyna}@ufal.mff.cuni.cz

**Abstract**

In this paper, we present a method of improving the accuracy of machine translation evaluation of Czech sentences. Given a reference sentence, our algorithm transforms it by targeted paraphrasing into a new synthetic reference sentence that is closer in wording to the machine translation output, but at the same time preserves the meaning of the original reference sentence. Grammatical correctness of the new reference sentence is provided by applying Depfix on newly created paraphrases. Depfix is a system for post-editing English-to-Czech machine translation outputs. We adjusted it to fix the errors in paraphrased sentences. Due to a noisy source of our paraphrases, we experiment with adding word alignment. However, the alignment reduces the number of paraphrases found and the best results were achieved by a simple greedy method with only one-word paraphrases thanks to their intensive filtering. BLEU scores computed using these new reference sentences show significantly higher correlation with human judgment than scores computed on the original reference sentences.

**Keywords:** paraphrases, machine translation, evaluation

## 1. Introduction

Metrics for automatic evaluation of machine translation (MT) are essential not only for measuring the quality of translations and comparing different systems and approaches, but also for the development of the translation systems themselves.

BLEU (Papineni et al., 2002) remains the most common metric for MT evaluation, even though other, better-performing metrics exist (Macháček and Bojar, 2013). BLEU is computed from the number of phrase overlaps between the translated sentence and the corresponding reference sentences, i.e., translations made by a professional human translator.

The advantage of BLEU is its simplicity and language independence. But it performs very badly for morphologically rich languages like the Czech language (Bojar et al., 2010), especially when only a single reference sentence is used.

One of the reasons is that BLEU disregards synonymous phrases and word form variants. One way how to alleviate this drawback is to include paraphrases to the evaluation metric (e.g. METEOR (Lavie and Agarwal, 2007)).

But this often awards even sentences with paraphrases that are not grammatically correct. We take a different approach by transforming a reference translation into a sentence that is closer to the MT output and keeps its original meaning and correctness.

We start with a basic algorithm for lexical (one-word) and phrase substitution based on phrasal alignment and tables of synonymous expressions. Further, we apply Depfix – a system originally designed for automatic correction of grammatical errors that appear often in English-to-Czech MT outputs, on newly created paraphrases.

Our method is independent on the evaluation metric. We use BLEU score in our experiments because of its common usage. Using our new reference sentences, BLEU achieves significant improvement of its correlation with human judgment.

## 2. Related Work

Targeted paraphrasing for MT evaluation is introduced in Kauchak and Barzilay (2006). They focus on lexical substitution in Chinese-to-English translations. They select all pairs of words for which one word appears in a reference sentence, second word in a hypothesis (the MT output), but none of them in both. They keep only pairs of synonymous words, i.e. words appearing in the same WordNet (Miller, 1995) synset. Each such a pair of words was further contextually evaluated. For every confirmed synonym, a new reference sentence is created by placing it to the reference sentence on the position of its synonym.

Our algorithm differs in many ways. As Czech belongs among inflective languages with rich morphology, a Czech word has typically many forms and the correct form depends heavily on its context, e.g., cases of nouns depend on verb valency frames. Therefore, we do not attempt to change a single word in a reference sentence but we focus on creating one single correct reference sentence.

Instead of the contextual evaluation, we focus on keeping grammatical correctness and the original meaning by using Depfix (Rosa et al., 2012) – an automatic post-editing system which is able to fix Czech sentences containing grammatical errors. Depfix was originally designed for post-editing outputs of English-to-Czech phrase-based machine translation. We adapted it to fit our setting.

Furthermore, as the Czech WordNet is substantially smaller, we exploit – in addition to this language source – another noisier source of paraphrases. Because of the noise, we experiment with adding alignment between the hypothesis and the corresponding reference sentence.

## 3. Data

### 3.1. Test Data

We use data sets from the English-to-Czech Translation Task of the Workshop on Statistical Machine Translation (WMT) from years 2012 and 2013 (Callison-Burch et al., 2012; Bojar et al., 2013). For WMT12, the data consists

| | WMT12 | WMT13 |
|---|---|---|
| WordNet | 780 | 650 |
| filtered Meteor | 4588 | 3877 |
| their union | 4766 | 4013 |

Table 1: Average number of pairs of words identified as a paraphrase between a MT output and a corresponding reference sentence according to their source.

of 13 files with (Czech) outputs of MT systems, one with corresponding reference sentences and one with original English source sentences. Each file contains 3003 sentences. The data for WMT13 contains 14[1] outputs of MT systems and similarly one file with Czech references and the original English source file. There are 3000 sentences in each file.

We perform morphological analysis and tagging of the MT outputs and the reference sentences using Morče (Spoustová et al., 2007).

### 3.2. Sources of Paraphrases

We use the following two available sources of Czech paraphrases.

#### 3.2.1. Czech WordNet 1.9 PDT

The first one is the Czech WordNet 1.9 PDT (Pala and Smrž, 2004). It is derived from the WordNet (Miller, 1995) by automatic translation followed by manual control. It contains rather high quality lemmatized paraphrases. On the other hand, their amount is insufficient for our purposes (see Table 1).

#### 3.2.2. Czech Meteor tables

Czech Meteor tables (Lavie and Agarwal, 2007) are an additional source of paraphrases. They are large in size, but they contain a lot of noise as they are constructed automatically from parallel data via pivoting (Bannard and Callison-Burch, 2005).

The noise was particularly high among the multiword paraphrases – for example: *svého názoru* (its opinion) and *šermovat rukama a mlátit neviditelného* (to flail one's arms and to beat the invisible one) are selected as a paraphrase.

Among one-word paraphrases the noise is sparser, but there are still pairs like *1873* - *pijavice* (a leech) or *afghánci* (Afghans) - *št'astně* (happily) identified as synonyms. However, the biggest problem is that most of synonymous pairs were just different word forms of the same lemma.

We therefore attempt to automatically filter the Meteor table, the methods are described in Section 5.

## 4. Algorithm

We experiment with several algorithms for paraphrasing reference sentences. They differ in the method for selecting potential paraphrase pairs and in the length of paraphrases.

[1] We use only 12 of them because two of them (FDA.2878 and online-G) have no human judgments of Czech system outputs.

### 4.1. Candidate Selection

We select potential paraphrases using two different methods. The first one is a simple greedy search similar to Kauchak and Barzilay (2006), the other one uses automatic word alignment for selecting corresponding segments of the reference sentence and the hypothesis.

#### 4.1.1. Simple Greedy Method

Let $W_L$, $R_L$ be sets of lemmas from the hypothesis (MT output) and the reference sentence, respectively. Then, one-word paraphrase candidates are chosen as:

$$C_L = \{(r,w) | r \in R_L \setminus W_L \wedge w \in W_L \setminus R_L \wedge r_{POS} = w_{POS}\}$$

Multi-words candidates $C_M$ are selected as the Cartesian product of all sequences from the reference sentence and all sequences from the hypothesis. Formally:

Let $r_1, ..., r_n, w_1, ..., w_m$ be the hypothesis and the reference sentence, respectively. Then the set of multi-word paraphrase candidates is selected the following way:

$$C_M = \{(< r_i, .., r_{i+x} >, < w_j, ..., w_{j+y} >) | \, 1 \leq i \leq n-x \wedge 1 \leq j \leq m-y \wedge 0 \leq x,y \leq 6 \wedge (x \neq 0 \vee y \neq 0)\}$$

Maximum phrase length is seven words, because that is the length of the longest paraphrases in the data.

#### 4.1.2. Word and Phrase Alignments

One possible way to make the algorithm more reliable is to restrict the application of paraphrases to words/phrases which are aligned to each other. We compute word alignment between the reference translation and MT system outputs using GIZA++ (Och and Ney, 2000).

If we used only our test data to create the alignment (13 x 3003 + 12 x 3000 = 75039 sentence pairs), the alignment quality would be insufficient. In order to make the training data for word alignment larger, we take advantage of the fact that all outputs are translations of the same data and also add all pairs of system outputs to our data, creating over 1,000,000 "artificial" sentence pairs. For example, the parallel data for WMT12 then looks as follows:

| Source | Target |
|---|---|
| system 1 | system 2 |
| system 1 | system 3 |
| ... | ... |
| system 1 | system 13 |
| system 1 | reference |
| system 2 | system 1 |
| system 2 | system 3 |
| ... | ... |
| system 13 | reference |

We also experiment with adding much larger synthetic parallel data created by machine translation (note that we need Czech-Czech data) but there was no impact on the quality of paraphrasing so we follow the outlined approach which requires no additional data or processing.

The set of one-word candidates $C_L$ is then simply the set of all word pairs such that there exists an alignment link

between them. The set $C_M$ is extracted using phrase extraction for phrase-based MT, the standard consistency criterion is applied (Och et al., 1999).

### 4.2. Paraphrasing

We reduce the set $C_L$ to pairs appearing in our paraphrase tables in the following way. If a word appears in several synonymous pairs we give preference to those found in WordNet or even better in the intersection of paraphrases from WordNet and filtered Meteor. Similarly, we filter $C_M$ to pairs also contained in the multi-word Meteor tables.
We evaluate three different paraphrasing methods which differ in the order of substitution.

**One-word only** We proceed word by word from the beginning of the reference sentence to its end. If a lemma of a word appears as the first member of a pair in reduced $C_L$, it is replaced by the word from hypothesis that has its lemma as the second element of that pair, i.e., paraphrase from the hypothesis. Otherwise, we keep the original word from the reference sentence.

**One-word first** We use *One-word only* and then we apply longer paraphrases. In that case we move ahead from the longest paraphrases to the shortest. That is because Meteor contains often even components of phrases and we could substitute, instead of whole phrase, only part of it. We do not attempt to replace any word that was already changed before.

**Multi-word first** We substitute the longest confirmed paraphrases from $C_M$ and move to the shorter ones. We replace again only sequences that have not been substituted yet. After this, we paraphrase the remaining unchanged words with the *One-word only* method.

### 4.3. Depfix

Depfix is an automatic post-editing system, originally designed for improving quality of phrase-based English-to-Czech machine translation outputs. It consists of a set of linguistically-motivated rules and a statistical component that correct various kinds of errors, especially in grammar (e.g. morphological agreement), using a range of natural language processing tools to provide analyses of the input sentences.
We observe that the errors that appear in the outputs of our paraphrasing algorithm are often similar to some errors appearing in outputs of phrase-based machine translation systems, e.g errors in morphological agreement are very common. This makes Depfix a good fit for fixing the errors, since typical grammar correcting tools, such as a grammar-checker in a word processor, focus on errors that are typical for humans, not for machines. For this reason, we apply Depfix post-editing to fix the errors in grammar that frequently appear in our outputs.
However, some error types that are common in phrase-based machine translation, such as errors in preserving the correct verb tense, do not frequently emerge in the paraphrasing process. Therefore, we experiment with two Depfix configurations in this work:

**full** the original Depfix system with all 33 fixing blocks, as described in (Rosa, 2013)

**limited** Depfix adapted for fixing paraphrasing errors by disabling 10 of the fixing blocks[2]

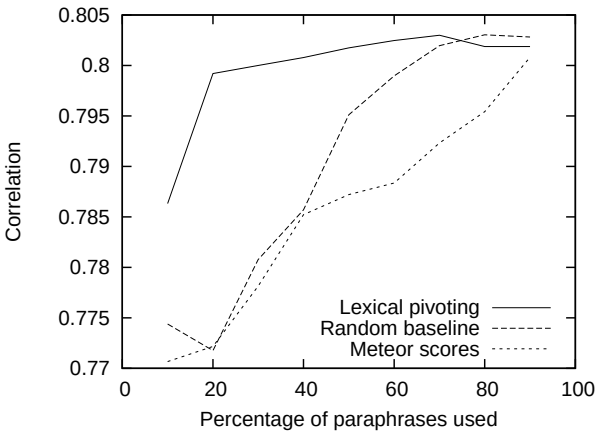


Figure 1: Comparison of automatic filtering techniques for *one-word* paraphrases on WMT12 data.

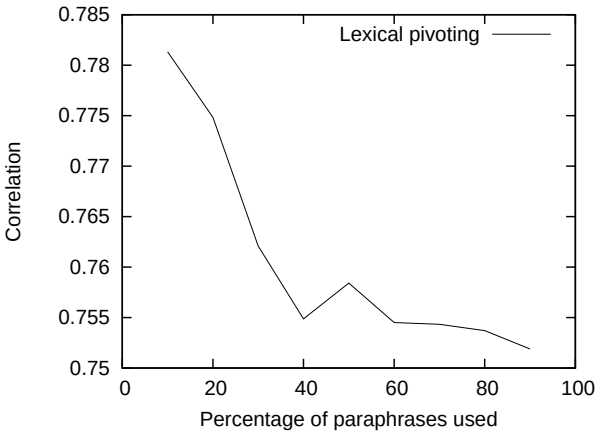


Figure 2: Automatic filtering of multi-word paraphrase for the *multi-word-first* scenario on WMT12 data.

## 5. Filtering the Meteor Tables

We try to remove the noise from the data with two different methods. The first one is based on manual error analysis and it is applied to one-word pairs only. The second one is fully automatic and can be applied on all data, but its results are inconclusive. Therefore, we employ only the first method in the rest of our experiments.

[2]The following fixing blocks are disabled in limited Depfix: Fixing reflexive tantum, Fixing morphological number of nouns, Translation of "by", Translation of "of", Translation of present continuous, Subject categories projection, Missing reflexive verbs, Subject personal pronouns dropping, Tense translation, Negation translation

### 5.1. Error-analysis Based Filtering

We manually examine sentences after paraphrasing using only Meteor tables. Based on our observation, we perform the following operations on pairs of one-word paraphrases from the Meteor tables:

- morphological analysis using Morče (Spoustová et al., 2007) and replacing of word forms with their lemmas;
- removing pairs of identical lemmas;
- removing pairs with different part of speech;
- removing pairs of unknown words (typically foreign words).

The last two rules have a single exception – paraphrases consisting of numeral and corresponding digits, e.g., *osmnáct* (eighteen) and *18*.[3] These paraphrases are very common in the data.

This way we reduce more than 160 000 pairs of one word paraphrases to only 32 154 couples of lemmas. All examples of bad one-word paraphrases from Subsection 3.2.2. are removed.

### 5.2. Automatic Filtering

Filtering strategies described in this section are based on assigning a score to each paraphrase pair. We then gradually remove paraphrases with low scores and measure the effect on the final correlation of our metric.

The first, straightforward approach is to use the paraphrase scores already provided in Meteor. They are based on phrasal translation probabilities and it corresponds to paraphrase probability in the pivoting model.

We propose an alternative scoring based on pivoting and lexical translation scores:

$$\text{lex\_p}(\mathbf{s},\mathbf{t}) = \sum_{s\in\mathbf{s}}\sum_{t\in\mathbf{t}}\sum_{pivot} \text{lex}(s|pivot)\text{lex}(pivot|t)$$

In this case, pivots are all words aligned to both $s$ and $t$ in the parallel data. To get lexical translation probabilities, we use maximum likelihood estimation from single best word alignment computed on CzEng 1.0 (Bojar et al., 2012). We refer to this score as *lexical pivoting*.

We use random selection as the baseline – paraphrases are simply shuffled and we then use the first 10, 20,... percent of them.

We only evaluate the filtering techniques on the WMT12 data. First, we attempt to filter one-word paraphrases and use the cleaner paraphrase table in the *one-word-only* paraphrasing strategy. Note that our paraphrase table has already been filtered using the error-analysis based filtering described above.

Figure 1 shows the performance of different filtering techniques for one-word paraphrases. Relying on Meteor scores proves worse than random selection. Using lexical pivoting, we can keep a high correlation even if we throw away as much as 90% of the paraphrases, however we do not improve (by a relevant margin) upon the baseline correlation of 0.802 achieved by *one-word-only* paraphrasing with the full paraphrase table.

We evaluate the best-performing technique also in the *multi-word-first* scenario where we use it for filtering multi-word paraphrases (see Figure 2). As we reduce the number of paraphrases, we observe a considerable improvement of correlation, however we never outperform *one-word-only* or *one-word-first*. In this case, the filtering simply mitigates the damage done by the multi-word paraphrases. We cannot hope to achieve a higher score without a more fine-grained grip on what a good multi-word paraphrase is.

[3] *0smnáct* has the part of speech *C*, which is designated for numerals, *18* is marked with *X* meaning it is an unknown word for the morphological analyzer.

| Method | Greedy selection | | Word alignment | |
|---|---|---|---|---|
| | Words | Phrases | Words | Phrases |
| One-word only | 1.59 | – | 0.86 | – |
| One-word first | 1.59 | 0.23 | 0.86 | 0.22 |
| Multi-word first | 1.38 | 0.31 | 0.81 | 0.27 |

Table 2: Average number of replaced words/phrases per sentence for each method on data from WMT12.

| Method | Greedy selection | | Word alignment | |
|---|---|---|---|---|
| | Words | Phrases | Words | Phrases |
| One-word only | 1.33 | – | 0.76 | – |
| One-word first | 1.33 | 0.20 | 0.76 | 0.20 |
| Multi-word first | 1.04 | 0.68 | 0.74 | 0.24 |

Table 3: Average number of replaced words/phrases per sentence for each method on data from WMT13.

## 6. Results

The performance of an evaluation metric in MT is usually computed as the Pearson correlation between the automatic metric and human judgment (Papineni et al., 2002). The correlation estimates the linear dependency between two sets of values. It ranges from -1 (perfect negative linear relationship) to 1 (perfect linear correlation).

The official manual evaluation metric of WMT12 (Callison-Burch et al., 2012) and WMT13 (Bojar et al., 2013) provides just a relative ranking: a human judge always compares the performance of five systems on a particular sentence. From these relative rankings, we compute the absolute performance of every system using the *greater than others* method, i.e., the score is based on how frequently the system is judged to be better than another system. Ties among several systems are ignored. We use this score as a human judgment in further evaluation.

Results of our method are presented in Tables 4 and 5. The baseline (i.e., using the original reference sentences) has a correlation of 0.749, 0.829 respectively. All evaluated approaches outperform it, the simplest one *One-word only* performs best (Figure 3 shows an example of this method). We use a freely available implementation[4] of Meng et al. (1992) to determine whether the difference in correlation

[4] `http://www.cnts.ua.ac.be/~vincent/scripts/rtest.py`

| Method | Greedy selection | | | Word alignment | | |
|---|---|---|---|---|---|---|
| | No Depfix | Full Depfix | Limited Depfix | No Depfix | Full Depfix | Limited Depfix |
| One-word only | 0.802 | 0.827 | **0.832** | 0.792 | 0.813 | 0.810 |
| One-word first | 0.785 | 0.822 | 0.816 | 0.767 | 0.792 | 0.798 |
| Multi-word first | 0.768 | 0.810 | 0.804 | 0.761 | 0.781 | 0.778 |

Baseline correlation: **0.749**

Table 4: Correlation of the human judgment and BLEU computed with the data from WMT12

| Method | Greedy selection | | | Word alignment | | |
|---|---|---|---|---|---|---|
| | No Depfix | Full Depfix | Limited Depfix | No Depfix | Full Depfix | Limited Depfix |
| One-word only | 0.861 | **0.887** | 0.883 | 0.856 | 0.877 | 0.872 |
| One-word first | 0.851 | 0.880 | 0.875 | 0.833 | 0.871 | 0.863 |
| Multi-word first | 0.838 | 0.870 | 0.864 | 0.833 | 0.868 | 0.861 |

Baseline correlation: **0.829**

Table 5: Correlation of the human judgment and BLEU computed with the data from WMT13.

| | |
|---|---|
| Source | *The location alone is classic.* |
| Hypothesis | *Samotné místo je klasické.*<br>Actual place is classic<br>The place alone is classic. |
| Reference | *Už poloha je klasická.*<br>Already position is classic.<br>The position itself is classic. |
| New reference | *Už místo je klasická.*<br>Already place is classic<br>*The place itself is classic. |
| Depfixed ref. | *Už místo je klasické.*<br>Already place is classic<br>The place itself is classic. |

Figure 3: Example of the *One-word only* method. The hypothesis is grammatically correct and has very similar meaning as the reference sentence. The new reference is closer in wording to the hypothesis, but there is no agreement between the noun and adjective. Depfix resolves the error and the final reference is correct and much similar to the hypothesis.

coefficients is statistically significant. The test shows that BLEU performs better with our reference sentences with 99% certainty.

Multi-word paraphrases are very noisy and while they do bring the system outputs closer to the reference (the average BLEU score of the systems increases), they often propose non-equivalent translations or violate the correctness of the sentence, thus blurring the differences between systems.

When paraphrasing is restricted by word alignment, all methods perform worse. As Tables 2 and 3 show, the number of applied paraphrases is much lower: while the proportion of correct paraphrases is higher, their amount is reduced too much and overall, our technique is harmed by this restriction.

On the other hand, applying Depfix is always beneficial, with the positive effects ranging from 0.017 up to 0.042. This supports our assumption of the importance of grammatical correctness of the created references. However, the *limited* version is not optimally chosen and performs worse than the *full* version in most cases.

Results on the data from WMT13 and WMT12 are very similar. Again, paraphrasing helps to increase the accuracy of the evaluation, even though the differences on the WMT13 data are not as big due to much higher baseline. This is also reflected in the smaller amount of substitutions (see Table 3).

## 7. Conclusion and Future Work

Our results confirm the positive impact of paraphrasing a reference sentence on the performance of the BLEU score. We evaluate a number of approaches to paraphrasing. The best results are achieved by the *one-word only* greedy substitution method. We achieve a statistically significant improvement in the evaluation of English-to-Czech MT.

We illustrate several methods for reducing noise in a paraphrase corpus and we confirm importance of grammar correctness of reference sentences in MT evaluation by the improvement of correlation after applying Depfix.

In the future, we plan to further increase the correlation by creating our own Czech paraphrase tables that would be larger than Czech WordNet, but less noisy than Czech Meteor Tables.

Another way to improve the performance of our system which we want to follow is a further adaptation of the Depfix system to our task. We intend to tune existing Depfix corrections, as well as to add new corrections specific to our task. We would also like to devise a way of informing Depfix which parts of the sentences come from the reference and which come from the paraphrasing to eliminate "false positives", i.e. Depfix attempting to correct words that are unlikely to be incorrect.

**Acknowledgment**

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

## 8. References

Bannard, Colin and Callison-Burch, Chris. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bojar, Ondřej, Kos, Kamil, and Mareček, David. (2010). Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 86–91, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Bojar, Ondřej, Buck, Christian, Callison-Burch, Chris, Federmann, Christian, Haddow, Barry, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu, and Specia, Lucia. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu, and Specia, Lucia. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. In *Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Kauchak, David and Barzilay, Regina. (2006). Paraphrasing for Automatic Evaluation. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 455–462, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lavie, Alon and Agarwal, Abhaya. (2007). Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Macháček, Matouš and Bojar, Ondřej. (2013). Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Meng, Xiao-Li, Rosenthal, Robert, and Rubin, Donald B. (1992). Comparing correlated correlation coefficients. *Psychological bulletin*, 111(1):172.

Miller, George A. (1995). WordNet: A Lexical Database for English. *COMMUNICATIONS OF THE ACM*, 38:39–41.

Och, Franz Josef and Ney, Hermann. (2000). Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

Och, Franz Josef, Tillmann, Christoph, and Ney, Hermann. (1999). Improved Alignment Models for Statistical Machine Translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28.

Pala, Karel and Smrž, Pavel. (2004). Building Czech WordNet. *In Romanian Journal of Information Science and Technology*, 7:79–88.

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rosa, Rudolf, Mareček, David, and Dušek, Ondřej. (2012). DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 362–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Spoustová, Drahomíra, Hajič, Jan, Votrubec, Jan, Krbec, Pavel, and Květoň, Pavel. (2007). 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.