# Targeted Paraphrasing of Czech Sentences for Machine Translation Evaluation

**Petra Barančíková**
Institute of Formal and Applied Linguistics
Charles University in Prague, Faculty of Mathematics and Physics
Malostranské náměstí 25, Prague, Czech Republic
barancikova@ufal.mff.cuni.cz

## Abstract

This thesis is focused on improving the accuracy of machine translation to Czech language using targeted paraphrasing. We develop and compare several approaches for creating new synthetic references that are closer in wording to a corresponding machine translations.

These new reference sentences make evaluation using traditional metrics more reliable as they lack of paraphrase and free word order support.

After implementation, the system for generating new synthetic references will be beneficial not only for the quality of MT evaluation, but also for improved tuning and development of the MT systems. This way, it will directly influence the quality of machine translation itself.

## 1 Introduction

Since the very first appearance of machine translation (MT) systems, a necessity for their objective evaluation and comparison has emerged. The traditional human evaluation while being the most reliable has serious drawbacks – it is time-consuming and expensive.

However, the main problem of human evaluation is that it is highly dependent on the person annotating and there is generally low inter-annotator agreement (Bojar et al., 2013a). Moreover, it is practically impossible to repeat the evaluation with the same results (Bojar, 2012).

Due to being slow and unreproducible, it is impossible to use human evaluation for tuning and development of MT systems. Well-performing automatic MT evaluation metrics are essential precisely for these tasks.

The pioneer metrics correlating well with human judgment were BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). They are computed based on n-gram overlap between the MT output (hypothesis) and one or more corresponding reference sentences, i.e., translations made by a human translator.

Due to its simplicity and language independence, BLEU still remains de facto the standard metric for MT evaluation and tuning, even though its correlation with human judgment is not as high as previously thought (Callison-Burch et al., 2006b) and other, better-performing metrics exist (Macháček and Bojar (2013), Bojar et al. (2014)).

Furthermore, obtaining reference sentences is labour intensive and expensive,[1] thus the standard practice is using only one reference sentence and BLEU then tends to perform badly.

As there are many translations of a single sentence, even a perfectly correct hypothesis might get a low score due to different wording and disregarding synonymous expressions (see Figure 1). This is especially valid for morphologically rich languages with free word order like the Czech language. (Bojar et al., 2010)

As the main task of an MT metric is essentially to decide whether a reference sentence is a paraphrase of a given hypothesis, our goal is to achieve higher accuracy of MT evaluation by targeted paraphrasing of reference sentences, i.e. creating a new synthetic reference sentence that is still correct and keeps the meaning of the original sentence but at the same time it is closer in wording to the MT output. BLEU and other string-based metrics performs more reliable using these new references.

The structure of the thesis is as follows: in the next section, we present other work in paraphrasing for MT evaluation. In Section 3, we introduced

[1] For example, production of reference translation at the Linguistic Data Consortium is complicated process involving translation by professional agencies based on elaborate guidelines and detailed quality control (Strassel et al., 2006).

| | | | | | | |
|---|---|---|---|---|---|---|
| Original sentence | *Banks are testing payment by mobile telephone* | | | | | |
| MT output | *Banky* | *zkoušejí* | *platbu* | *pomocí* | *mobilního* | *telefonu* |
| | Banks | are testing | payment | with help | mobile | phone |
| | Banks are testing payment by mobile phone | | | | | |
| Reference sentence | *Banky* | *testují* | *placení* | *mobilem* | | |
| | Banks | are testing | paying | by mobile phone | | |
| | Banks are testing paying by mobile phone | | | | | |

Figure 1: Example from WMT12 - Even though the translation is grammatically correct and the meaning of both sentences is very similar, it doesn't contribute to the BLEU score. There is only one unigram overlapping.

the data we use in our experiments. In sections 4 and 5, we show paraphrasing based on phrase substitutions and machine translation itself. Finally, we conclude with avenues for further work.

## 2 Related Work

There are two attitudes towards solving the previously described problem - one is to change the automatic metric itself to be tolerant to other representations of a sentence and the other one is to preprocess automatically the reference sentence via paraphrasing.

### 2.1 Automatic Metrics

Second generation metrics Meteor (Denkowski and Lavie, 2014), TERp (Snover et al., 2009) and ParaEval (Zhou et al., 2006) still largely focus on a n-gram overlap while including other linguistically motivated resources. They utilize paraphrase support using their own paraphrase tables and show higher correlation with human judgment than BLEU.

Meteor is available for several languages including Czech. It explicitly addresses weaknesses of BLEU – it takes into account recall, distinguishes between functional and content words, allows language-specific tuning of parameters and many others. The standard setting of Meteor for evaluation of Czech sentences offers two levels of matches – exact and paraphrase. As we show further (see Section 4.5), the Czech Meteor paraphrase tables are so noisy that they actually harm the performance of the metric, as it can award mistranslated and even untranslated words.

String matching is hardly discriminative enough to reflect the human perception and there is growing number of metrics that compute their score based on rich linguistic features – matching based on parse trees, POS tags or textual entailment (e.g. Liu and Gildea (2005), Owczarzak et al. (2007), Amigó et al. (2009), Padó et al. (2009), Macháček and Bojar (2011)).

These metrics shows better correlation with human judgment, but their wide usage is limited by being complex and language-dependent. As a result, there is a trade-off between linguistic-rich strategy for better performance and wide applicability of simple string level matching.

### 2.2 Sentence paraphrasing

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, 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 is further contextually evaluated. For every confirmed word, a new reference sentence is created by placing it to the reference sentence on the position of its synonym.

This solution is not directly applicable to the Czech language. As Czech belongs to inflective languages with rich morphology, a Czech word has typically many forms and the correct form depends heavily on its context, e.g., morphological cases of nouns depend on verb valency frames. Changing a single word may result into not grammatical sentence. Therefore, we do not attempt to change a single word in a reference sentence but we focus on creating one single correct reference sentence.

There are many other applications for paraphrases in the MT field. Paraphrases are utilized for translating out-of-vocabulary words and increasing quality of MT systems trained on sparse data (Callison-Burch et al., 2006a), (Marton et al., 2009) or enlarge training data (Nakov, 2008). Madnani (2010) and Mehay and White (2012) show that single translation augmented by targeted paraphrases can successfully replace up to four additional human reference translations in the tuning phase of a MT system.

## 3 Data

### 3.1 Text Data

We perform our experiments on data sets from the English-to-Czech translation task of WMT12 (Callison-Burch et al., 2012), WMT13 (Bojar et al., 2013a). The data sets contain 13/14[2] files with Czech outputs of different MT systems. Each file contains 3003/3000 sentences.

In addition, each data set contains one file with corresponding reference sentences and one with original English source sentences. We perform morphological analysis and tagging of the hypotheses and the reference sentences using Morče (Spoustová et al., 2007).

The human judgment of hypotheses is available as a relative ranking of performance of five systems for a sentence. We calculated the absolute performance of every system by the "> others" method (Bojar et al., 2011), which was the WMT12 official system score. It is computed as $\frac{wins}{wins+loses}$, ties among several systems are ignored. We use this score as a human judgment in further evaluation.

### 3.2 Sources of Paraphrases

We make use of two existing sources of Czech paraphrases – Czech WordNet 1.9 PDT (Pala et al., 2011) and the Meteor paraphrase tables (Denkowski and Lavie, 2010).

**Czech WordNet 1.9 PDT** is derived from WordNet (Miller, 1995) by automatic translation followed by manual verification. It contains rather high quality lemmatized paraphrases. However, their amount is insufficient for our purposes - it contains only 13k pairs of synonymous lemmas.

On the other hand, Czech **Meteor paraphrase tables** are quite the opposite of Czech WordNet – they are large in size, but contain a lot of noise, as they were constructed automatically via *pivoting* (Bannard and Callison-Burch, 2005). The pivot method is an inexpensive way of acquiring paraphrases from large parallel corpora. It is based on the assumption that two phrases that share a meaning may have a same translation in a foreign language (Dyvik, 1998).

The noise is 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. We experimented with several methods of reducing the noise in the multi-word paraphrases, however, none of them outperforms omitting them completely (Barančíková et al., 2014).

Among one-word paraphrases the noise is sparser, but the paraphrase tables still contains pairs such as *pijavice* ("a leech") - *1873* or *afghánci* ("Afghans") - *šťastně* ("happily"). A lot of synonymous pairs are different word forms of a same lemma.

We perform automatic filtering among the one-word paraphrases in the Meteor table in the following way. Using Morče, we first perform morphological analysis of all one-word pairs and replace the word forms with their lemmas. We keep only pairs of different lemmas. Further, we dispose of pairs of words that differ in their parts of speech[3] (POS) or contain an unknown (typically foreign) word.

In this way we have reduced 684k paraphrases in the original Czech Meteor Paraphrase tables to only 32k pairs of lemmas. We refer to this paraphrase table as to *filtered Meteor*.

## 4 Simple substitution paraphrasing

In this section, we experiment with several algorithms for paraphrasing reference sentences widely based on Kauchak and Barzilay (2006). First, we select candidates for paraphrasing between a hypothesis and its corresponding reference sentence. For one-word paraphrases we search between different lemmas as the filtered Meteor and Czech WordNet are lemmatized, longer paraphrases are selected in their word forms.

After filtering according to a particular paraphrase table, synonymous expressions are substi-

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

[3] With a single exception – paraphrases consisting of numeral and corresponding digits, e.g., *osmnáct* ("eighteen") and *18* - *osmnáct* 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.

tuted from a hypotheses to a reference sentence.

The newly created reference may be non grammatical due to the direct substitution. To fix it, we employ Depfix (Rosa et al., 2012), a system for automatic correction of grammatical errors.

### 4.1 Candidate Selection

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

#### Simple Greedy Method

Let $w_1, ..., w_m, r_1, ..., r_n$ be the hypothesis and the reference sentence, respectively. We performed their tagging and extracted sets of lemmas $W_L$, $R_L$. Then, one-word paraphrase candidates are chosen as:

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

Multi-words candidates $C_M$ are selected analogically from all sequences of words from the reference sentence and from the hypothesis. Maximum phrase length is seven words (as is the length of the longest paraphrases in the data).

#### Word and Phrase Alignments

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

However, 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 $C_M$ to pairs contained in the multi-word Meteor tables and $C_L$ to Czech WordNet and the filtered Meteor. If there is a lemma contained in several pairs in $C_L$, we give preference to those found in WordNet or even better in the intersection of paraphrases from WordNet and filtered Meteor.

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. We move ahead from the longest paraphrases to the shortest. That is because the Meteor paraphrase tables contain 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.

### 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 remaining unchanged words with the *One-word only* method.

## 4.3 Depfix

As we substitute a word form directly from a hypothesis, it may happen that a resulting new reference is not grammatically correct. We rectify these errors by applying an automatic post-editing system Depfix (Rosa et al., 2012).

Depfix was originally designed for post-editing outputs of English-to-Czech phrase-based machine translation. 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 analyse of the input sentences.

We observe that the errors appearing 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 an appropriate tool for fixing the errors, since typical grammar correcting tools, such as a grammar-checker, focus on errors that are typical for humans, not for machines.

## 4.4 Results

Results of our method are presented in Table 1 as the Pearson correlation between human judgment and BLEU computed on our new references. All evaluated approaches outperform the baseline (i.e., using the original reference sentences), the simplest one *One-word only* performs best (Figure 2 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 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 is higher), they often propose non-equivalent translations or violate the correctness of the sentence, thus blurring the differences between systems.

| 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 ref. | *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 2: 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 more similar to the hypothesis.

When paraphrasing is restricted by word alignment, all methods perform worse. As Table 2 shows, 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 0.021 up to 0.058. This supports our assumption of the importance of grammatical correctness of the created references.

Results on the data from WMT12 and WMT13

**WMT12**

| 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 |

**WMT13**

| 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 2: Average number of replaced words/phrases per sentence.

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

**WMT12**

| Method | Greedy selection | | Word alignment | |
|---|---|---|---|---|
| | No Depfix | After Depfix | No Depfix | After Depfix |
| One-word only | 0.804 | **0.834** | 0.794 | 0.815 |
| One-word first | 0.788 | 0.825 | 0.763 | 0.800 |
| Multi-word first | 0.755 | 0.813 | 0.753 | 0.795 |

Baseline correlation: **0.751**

**WMT13**

| Method | Greedy selection | | Word alignment | |
|---|---|---|---|---|
| | No Depfix | After Depfix | No Depfix | After Depfix |
| One-word only | 0.865 | **0.891** | 0.860 | 0.881 |
| One-word first | 0.854 | 0.884 | 0.836 | 0.874 |
| Multi-word first | 0.841 | 0.875 | 0.840 | 0.871 |

Baseline correlation: **0.834**

Table 1: Correlation of the human judgment and BLEU on new references created by the simple substitution method.

are very similar – 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.

### 4.5 Meteor without paraphrase support

Based on the positive impact of filtering Meteor paraphrase tables for targeted lexical paraphrasing of reference sentences, we experiment with filtering them yet again, but this time as an inner part of the Meteor evaluation metric (i.e. for the paraphrase match).

The filtering of the paraphrase tables is performed analogically. We experiment with six different settings that are presented in Table 3. All of them are created by reducing the original Meteor paraphrase tables, except for the setting referred to as **WordNet**, which has its paraphrase table generated from one-word paraphrases in Czech WordNet to all their possible word forms appearing in CzEng (Bojar et al., 2012).

The results of our experiments are presented in Table 4. They are very consistent for WMT12 and WMT13. We show that independently of a reference sentence used, reducing the Meteor paraphrase tables in evaluation is always beneficial. The Meteor metric with exact match only on paraphrased references significantly outperforms Meteor with paraphrase support on original references.

**Different Lemma** and **WordNet** settings give the best results on the original reference sentences. That is because they are basically a limited version of the paraphrase tables we use for creating our new references, which contain both all different lemmas of the same part of speech from the Meteor paraphrase tables and all lemmas from Czech WordNet.

The main reason of the worse performance of the metric when employing the Meteor paraphrase tables is the noise. The metric may award even parts of the hypothesis left untranslated, as the original Meteor paraphrase tables contain some English words and their Czech translations as paraphrases, for example: *pšenice - wheat*[5], *vůdce - leader*, *vařit - cook*, *poloostrov - peninsula*.

## 5 Paraphrasing using Machine Translation

While the one-word substitution method offers good results, it is very limited. We would like to be able to use longer paraphrases, word order changes, switch between active and passive construction, etc.

For this purpose, we employ machine translation itself. There are many tools for MT and there is a close resemblance between translation and paraphrasing. They both attempt to preserve the meaning of a sentence, the first one between two

[5] In all examples the Czech word is the correct translation of the English side.

| setting | size | description of the paraphrase table |
|---:|:---:|:---|
| **Standard** | 684k | The original Meteor paraphrase tables |
| **One-word** | 181k | **Standard** without multi-word pairs |
| **Same POS** | 122k | **One-word** + only same part-of-speech pairs |
| **Different Lemma** | 71k | **Same POS** + only forms of different lemma |
| **Exact match** | 0 | No paraphrase tables |
| **WordNet** | 202k | Paraphrase tables generated from Czech WordNet |

Table 3: Different paraphrase tables for Meteor and their size (number of paraphrase pairs).

**WMT12**

| references | Standard | One-word | Same POS | Different Lemma | Exact match | WordNet |
|---|---|---|---|---|---|---|
| Original references | 0.833 | 0.836 | 0.840 | 0.863 | 0.861 | 0.863 |
| Before Depfix | 0.905 | 0.908 | 0.911 | 0.931 | 0.931 | 0.931 |
| New references | 0.927 | 0.930 | 0.931 | 0.950 | **0.951** | **0.951** |

**WMT13**

| references | Standard | One-word | Same POS | Different Lemma | Exact match | WordNet |
|---|---|---|---|---|---|---|
| Original references | 0.817 | 0.820 | 0.823 | 0.850 | 0.848 | 0.850 |
| Before Depfix | 0.865 | 0.867 | 0.869 | 0.895 | 0.895 | 0.894 |
| New references | 0.891 | 0.892 | 0.893 | **0.915** | **0.915** | **0.915** |

Table 4: Pearson's correlation of Meteor and the human judgment on original reference sentences and sentences created by the *One-word only* method.

languages and the second one within a single language by different word choice. It seems only natural to attempt to adjust some MT tools to translate within a single language for targeted paraphrasing.

We describe this attempt on two types of MT systems – phrase-based and rule-based. Initially, we experiment with the freely available SMT system Moses (Koehn et al., 2007). However, the results of this method are inconclusive. In the view of errors appearing in the new paraphrased sentences, we propose another solution – targeted paraphrasing using a rule-based translation system TectoMT (Žabokrtský et al., 2008) included in the NLP framework Treex (Popel and Žabokrtský, 2010).

### 5.1 Paraphrasing via Moses

Moses is a freely available statistical machine translation engine. In a nutshell, statistical machine translation involves the following phases: creating language and translation models, parameter tuning and decoding. We use Moses in the phrase-based setting. Simple scheme of Moses is presented in Figure 3.

A language model is responsible for a correct word order and grammatical correctness of the translated sentence. A translation model (a phrase table) supplies all possible translations of a word or a phrase or in our case, all possible paraphrases. Models are assigned weights which are learned during the parameter tuning phase.

During the decoding phase, all these models are combined to maximize $\sum_i \lambda_i \phi_i(\bar{f}, \bar{e})$, where $\lambda_i$ is a weight of the sub-model $\phi_i$ and $\bar{f}, \bar{e}$ is a hypothesis and a source sentence, respectively. In our case, we want to make a reference sentence closer to a corresponding machine translation output – $\bar{e}$ is the reference sentence and $\bar{f}$ is a new synthetic reference.

Moses with this setting could create paraphrases, but they would be just random paraphrases of the reference sentence – their similarity to our original hypotheses would not be guaranteed. Therefore, we also add a new feature for targeted paraphrasing to Moses.

#### 5.1.1 Language model

We create the language model (LM) using SRILM (Stolcke, 2002), a toolkit for building and applying statistical language models, on the data from the Czech part of the Czech-English parallel corpus CzEng (Bojar et al., 2012).

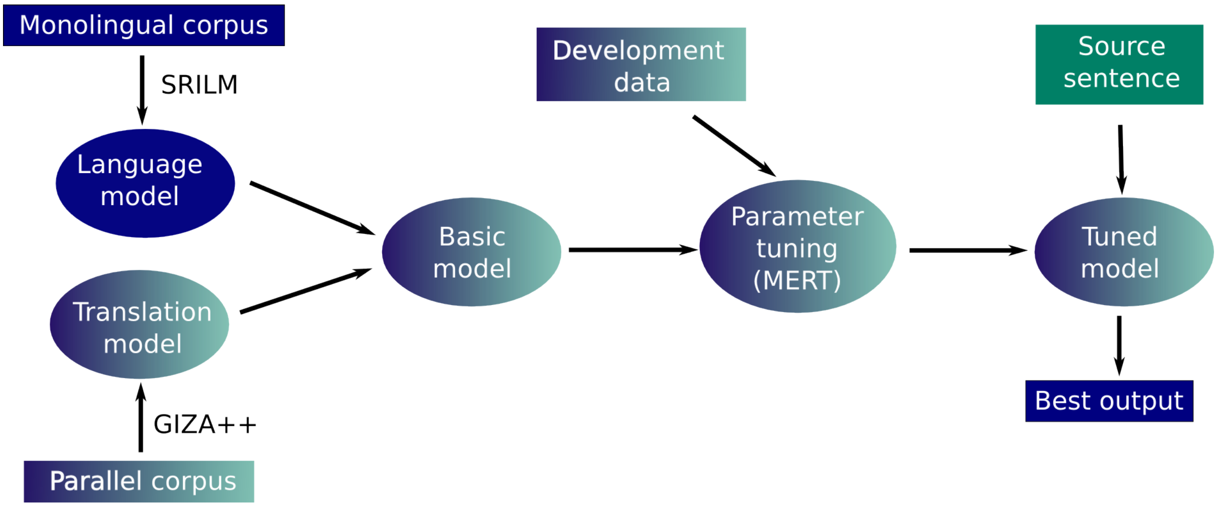
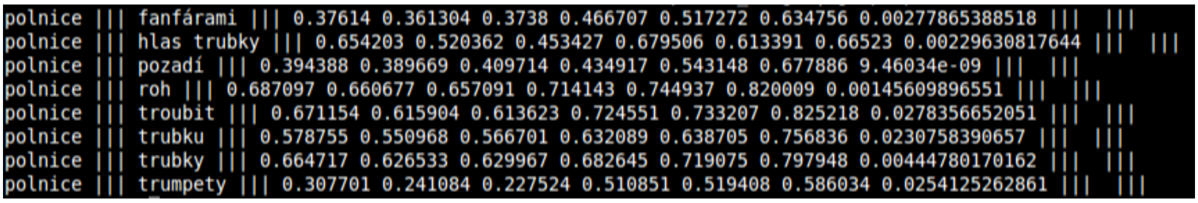


Figure 3: Simplified scheme of the translation system Moses. The blue colour represents a target language and the green colour represents a source language.



```
polnice ||| fanfárami ||| 0.37614 0.361304 0.3738 0.466707 0.517272 0.634756 0.00277865388518 |||  |||
polnice ||| hlas trubky ||| 0.654203 0.520362 0.453427 0.679506 0.613391 0.66523 0.00229630817644 |||  |||
polnice ||| pozadí ||| 0.394388 0.389669 0.409714 0.434917 0.543148 0.677886 9.46034e-09 |||  |||
polnice ||| roh ||| 0.687097 0.660677 0.657091 0.714143 0.744937 0.820009 0.00145609896551 |||  |||
polnice ||| troubit ||| 0.671154 0.615904 0.613623 0.724551 0.733207 0.825218 0.0278356652051 |||  |||
polnice ||| trubku ||| 0.578755 0.550968 0.566701 0.632089 0.638705 0.756836 0.0230758390657 |||  |||
polnice ||| trubky ||| 0.664717 0.626533 0.629967 0.682645 0.719075 0.797948 0.00444780170162 |||  |||
polnice ||| trumpety ||| 0.307701 0.241084 0.227524 0.510851 0.519408 0.586034 0.0254125262861 |||  |||
```

Figure 4: Excerpt from the Enhanced Meteor table.

### 5.1.2 Translation models

Each entry in Moses phrase tables contains a phrase, its translation, several feature scores (translation probability, lexical weight etc.), and optionally also alignment within the phrase and frequencies of phrases in the training data. The phrase tables are learned automatically from large parallel data. As we do not have any large corpora of Czech-Czech parallel data, we create the following two “fake” translation tables for paraphrasing from Czech WordNet and the Meteor paraphrase tables.

#### Enhanced Meteor table

The enhanced Meteor table was created from the Czech Meteor paraphrase tables. Each paraphrase pair comes with a pivoting score (see Section 3) which we adapt as a feature in out phrase table.

We also add our own paraphrase scores, acquired by *distributional semantics*. Distributional semantics assumes that two phrases are semantically similar if their contextual representations are similar (Miller and Charles, 1991).

We collect all contexts (words in a window of limited size) in which Meteor paraphrases occur in the Czech National Corpus (Křen et al., 2010) and then measure context similarity cosine distance, taking into account the number of word occurrences for each pair of paraphrases.

We add six scores for each pair of paraphrases according to the size of the context window used (1-3 words) and whether word order played a role in the context. Several lines from the Enhanced Meteor table are presented in Figure 4.

#### One-word paraphrase table

We create second phrase table from Czech WordNet and the filtered Meteor. It contains only one-word pairs and thus decreases the advantage of using a phrase translation system. However, it is designed to compensate for the noise in the Enhanced Meteor table (see Section 3).

We first create a set of all words from Czech side of CzEng appearing at least five times to exclude rare words and possible typos. We also add all words appearing in the MT outputs and the reference sentences. Morphological analysis of the words was then performed using Morče (Spoustová et al., 2007).

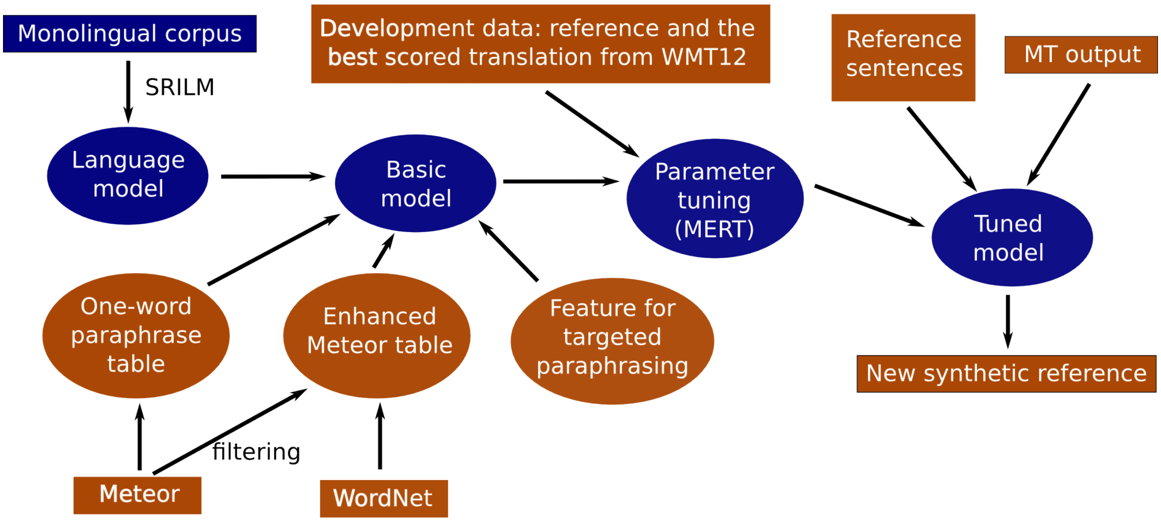


Figure 5: Pipeline of Moses adjusted for the targeted monolingual translation. Changes to the original Moses pipeline (Figure 3) are highlighted by the brown colour.

For every word $x$ from this set, we add to this translation table every pair of words that fulfils at least one of the following requirements:

- $x, x$ (not every word should be paraphrased)
- $x, y$, if $x$ has the same lemma as $y$ (some word might have different morphology in the paraphrased sentence)
- $x, y$, if lemma of $x$ and lemma of $y$ are paraphrases according to Czech WordNet PDT 1.9.
- $x, y$, if lemma of $x$ and lemma of $y$ are paraphrases according to the filtered Meteor.

These categories constitute the first four scores in the phrase table. A pair of words gets score $e$ if these words fall to a given category, 1 ($e^0$) otherwise.[6] This phrase table contains more than a million pairs of words.

We add another score expressing POS tag similarity between the two words. It is computed $e^{\frac{1}{a+1}}$, where $a$ is the minimal Hamming distance between tags of the words. This probability should reflect how morphologically distant the paraphrases are.

### 5.1.3 Feature for targeted paraphrasing

In order to steer the MT decoder (translation engine) in the direction of the hypotheses, we implemented to Moses an additional feature, which measures the overlap with the hypothesis. In order to keep its computation tractable during search, the overlap is defined simply as the number of words from the hypothesis confirmed by the reference translation. Our code is included in Moses.[7]

Schema of Moses modified for paraphrasing is presented in Figure 5.

### 5.1.4 Parameter tuning

We use the minimum error rate training (MERT) (Och, 2003) to find the optimal weights for our models. MERT asserts the weights to maximize the translation quality, which is measured with BLEU. We use the reference sentences and the highest rated MT output from WMT12 as the parallel development data for tuning.

This method, however, turned out not to be optimal for our setting. The feature for targeted paraphrasing naturally obtains the highest weight (0.51) as it provides an oracle guide towards the hypothesis. On the other hand, the language model get very small weight (0.016).

As a result, the paraphrased sentences tend to be closer to the hypothesis, but not grammatically correct. Therefore, we experiment with increasing the weight of the language model manually.

### 5.1.5 Results

We compare four different basic settings, the results are presented in Table 5. In contrast to our

[6] Phrase-table scores are considered log-probabilities.

[7] https://github.com/moses-smt/mosesdecoder/blob/master/moses/FF/CoveredReferenceFeature.cpp

| setting | reference sentence used | correlation | avg. BLEU |
|---|---|---|---|
| **Baseline** | original reference sentence, no paraphrasing | **0.75** | 12.8 |
| **Paraphrased** | paraphrased by Moses using MERT-learned weights | 0.50 | 15.8 |
| **LM+0.2** | paraphrased by Moses with LM weight increased by 0.2 | 0.24 | 9.1 |
| **LM+0.4** | paraphrased by Moses with LM weight increased by 0.4 | 0.22 | 6.7 |

Table 5: Description of the basic settings and the results - Pearson's correlation of BLEU and the human judgment, the average BLEU scores.

| **Source** | *Paclík claims he would dare to manage the association.* |
|---|---|
| **Baseline** | Paclík tvrdí , že by si na vedení asociace troufl.<br>*Paclík claims he would dare to lead the association.* |
| **Hypothesis** | Paclík tvrdí, že by se odvážil k řízení komory.<br>*Paclík claims he would find the courage to control the chamber.* |
| **Paraphrased** | Paclík tvrdí, že by se na řízení organizace troufl.<br>**Paclík claims he would dare to control the organization.* |
| **LM+0.2** | Paclík tvrdí, že by si troufl na řízení ekonomiky.<br>*Paclík claims he would dare to control the economy.* |
| **LM+0.4** | Říká se, že Paclík si troufl na řídící rady.<br>*They say that Paclík ventured to governing boards.* |

Figure 6: Example of the targeted paraphrasing using Moses. The hypothesis is a correct translation of the source sentence. The new paraphrased reference is slightly closer in wording to the hypothesis, but there is an error due to a bad word choice. The sentences created with increased weights of the language model are both grammatically correct, but the sentence lost its original meaning. In the **LM+0.4** setting, they also differ a lot in wording from both the hypothesis and the reference sentence.

previous results, the baseline score is not exceeded by any of our paraphrasing methods. Figure 6 represents an example of outputs.

There are several reasons for the clear decrease in correlation with paraphrased references. Hypotheses generated by the **Paraphrased** setting, while obtaining a substantially higher BLEU score, were mostly ungrammatical and reduced the correlation of our metric.

The small weight of the language model seems to be the problem, but its increase brings even more chaos. It creates hypotheses which are nice and grammatically correct but often wholly unrelated to the source sentence.

This shows that our paraphrase table noise filtering was by no means sufficient and there is a lot of noise in our phrase tables – given the high weight for the targeted paraphrase feature, we essentially transform the correct reference sentences to incorrect hypotheses at all cost, using our noisy phrase tables.

Our targeting feature is not ideal – it ignores word order and operates only on the word level (it does not model phrases). Ungrammatical translations with scrambled word order are considered perfectly fine as long as the translation contains the same words as the reference. So while the feature for targeted paraphrasing does provide a kind of oracle, it does not guarantee reaching the best possible translation in terms of BLEU score, let alone a grammatical translation.

Another problem is illustrated by very small weights assigned to our translation models. In fact, the highest weight (0.031) among translation model features was assigned to the tag similarity feature. This shows that our model features (pivoting score, distributional similarity scores, ...) fail to distinguish good paraphrases from the noise.

The combination of noise in the translation tables and the boosted language model then causes that the most common paraphrase according to the language model with a similar tag gets the preference.

### 5.2 Paraphrasing via Treex

Based on the previous results, Moses does not seem to be an optimal tool for our task, unless we create less noisy phrase tables and a better targeting feature, and unless we employ another func-

| | |
|---|---|
| Source | *The Internet has caused a boom in these speculations.* |
| Hypothesis | Internet vyvolal boom v těchto spekulacích .<br>*Internet caused boom in these speculations .*<br>*The Internet has caused a boom in these speculations.* |
| Reference | Rozkvět těchto spekulací způsobil internet .<br>*Boom these speculations caused internet .*<br>*Boom of these speculation was caused by the Internet.* |

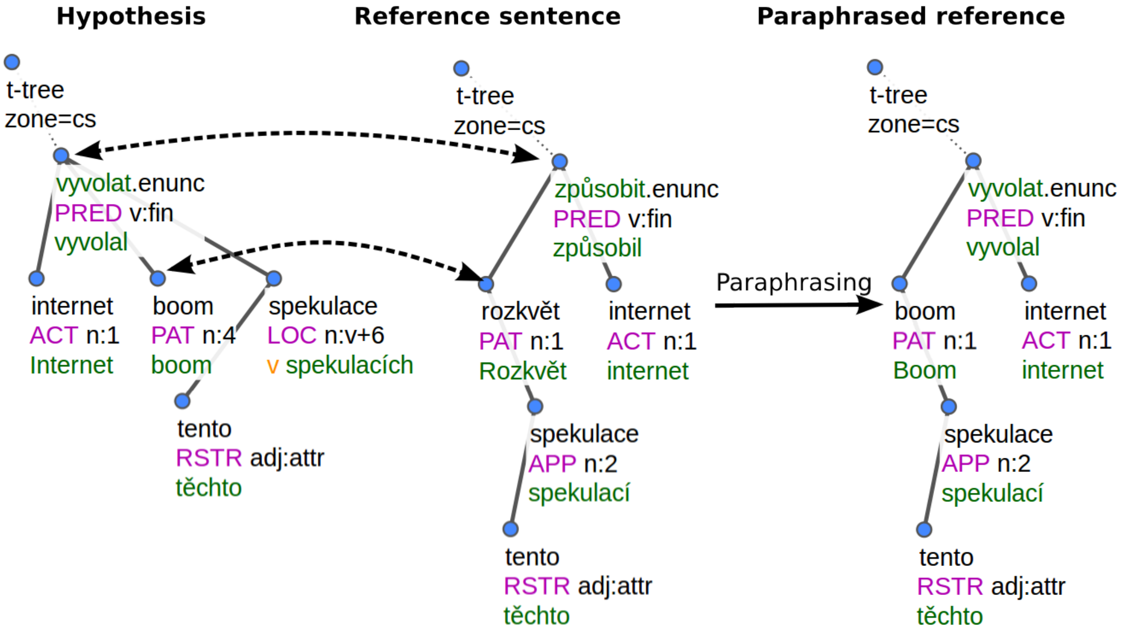


Figure 7: Example of the paraphrasing. The hypothesis is grammatically correct and has very similar meaning as the reference sentence. We analyse both sentences to t-layer, where we create new reference sentence by substituting synonyms from hypothesis to the reference. In the next step, we will change also the word order to better reflect the hypothesis.

tion for tuning weights.

However, there is another solution to a phrase-based translation system – namely a rule-based machine translation system TectoMT (Žabokrtský et al., 2008), which is included in a highly modular NLP software system Treex (Popel and Žabokrtský, 2010).

Treex implements the stratificational approach to language, adopted from the Functional Generative Description theory (Sgall, 1967) and its later extension applied in the Prague Dependency Treebank (Bejček et al., 2013). It represents sentences at four layers:

- **w-layer:** word layer; no linguistic annotation
- **m-layer:** morphological layer; sequence of tagged and lemmatized words
- **a-layer:** shallow-syntax/analytical layer; sentence is represented as a surface syntactic dependency tree
- **t-layer:** deep-syntax/tectogrammatical layer; sentence is represented as a dependency tree, where autosemantic words only have their own nodes; t-nodes consist of a t-lemma and a set of attributes – a *formeme* (information about the original syntactic form) and a set of *grammatemes* (essential morphological features).

In TectoMT, a sentence in a source language is analyses from the w-layer to the t-layer, where is transferred to the t-layer of a target language, and then generated to the w-layer of the target language.

Our analysis and generation pipeline is taken from the TectoTM system. In our setting, we transfer a hypothesis and a corresponding reference sentence to the t-layer, where we replace the transfer phase with a module for t-lemma paraphrasing. After paraphrasing, we perform synthesis to a-layer, where we plug in a reordering module and continue with synthesis to the w-layer.

This way, we can easily overcome some of the problems of paraphrasing using Moses. Most importantly, we can compare two sentences only and there is no need to create translation tables, thus less space for the noise to interfere. Also there already is highly developed machinery to avoid ungrammatical sentences.

Treex is opensource and is available on GitHub[8], including our modifications.

### 5.2.1 Analysis from w-layer to t-layer

The analysis from the w-layer to the a-layer includes tokenization, POS-tagging and lemmatization using MorphoDiTa (Straková et al., 2014), dependency parsing using the MSTParser (McDonald et al., 2005) adapted by Novák and Žabokrtský (2007), trained on PDT.

A surface-syntax a-tree is then converted into a deep-syntax t-tree. Auxiliary words are removed, with their function now represented using t-node attributes (grammatemes and formemes) of autosemantic words that they belong to, e.g. two a-nodes of the verb form *spal jsem* ("I slept") would be collapsed into one t-node *spát* ("sleep") with the tense grammateme set to past; *v květnu* ("in May") would be collapsed into *květen* ("May") with the formeme *v+X* ("in+X").

We choose the t-layer for paraphrasing, because the words from the sentence are lemmatized with their syntactical information hidden in formemes. Furthermore, functional words, which we do not want to paraphrase and that cause a lot of noise in our paraphrase tables, do not appear here.

### 5.2.2 Paraphrasing

The paraphrasing module T2T::ParaphraseSimple is available on GitHub.[9]

T-lemma of a reference t-node is changed from A to B if and only if:

1. there is no hypothesis t-node with lemma A
2. there is a hypothesis t-node with lemma B
3. there is no reference t-node with lemma B
4. A and B are paraphrases according to our paraphrase tables

The other attributes of the t-node are kept unchanged based on the theory that semantic properties are independent of the t-lemma. However, in practice, this is not always true: t-nodes corresponding to nouns are marked for grammatical gender, which is very often a grammatical property of the given lemma with no effect on the meaning (for example, "a house" can be translated either as a masculine noun *dům* or as feminine noun *budova*).

Therefore, when paraphrasing a t-node that corresponds to a noun, we delete the value of the gender grammateme, and let the subsequent synthesis pipeline generate the correct value of the morphological gender feature value (which is necessary to ensure correct morphological agreement with surrounding words, such as adjectives and verbs).

### 5.2.3 Synthesis from t-layer to a-layer

In this phase, a-nodes corresponding to auxiliary words and punctuation are generated, morphological feature values on a-nodes are initialized and set to enforce a morphological agreement among the nodes. Correct word forms based on lemmas and POS, and morphological features are generated using MorphoDiTa.

### 5.2.4 Tree-based reordering

The reordering block A2A::ReorderByLemmas is available on GitHub.[10]

The idea behind the block is to make the word order of a new reference as similar to the word order of the translation as possible, but with some tree-based constraints to avoid ungrammatical sentences.

The general approach is to reorder the subtrees rooted at modifier nodes of a given head node so that they follow in an order that is on average similar to their order in the translation. Figure 8 shows the reordering process of the a-tree from Figure 7.

Our reordering proceeds in several steps. Each a-node has an order, i.e. its position in the sentence. We define the *MT order* of a reference a-node as the order of its corresponding hypothesis a-node, i.e. a node with the same lemma.

[8] `https://github.com/ufal/treex`

[9] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/ParaphraseSimple.pm`

[10] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/A2A/ReorderByLemmas.pm`

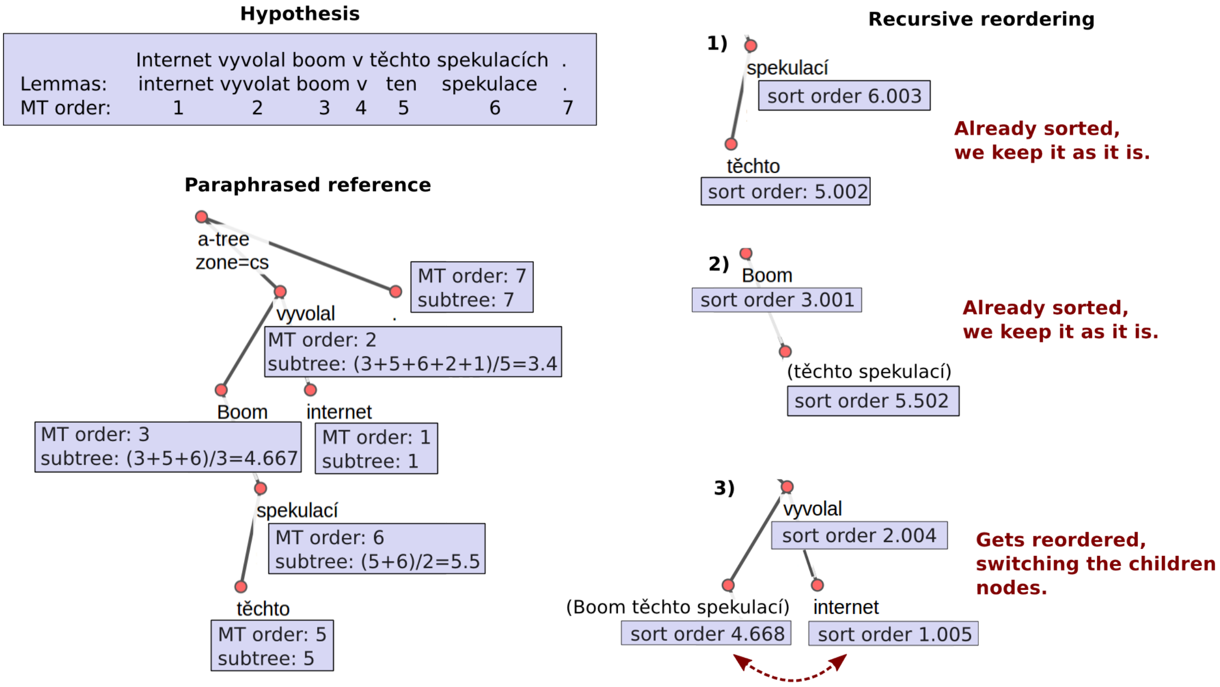


Figure 8: Continuation of Figure 7, reordering of the paraphrased reference sentence.

We set the MT order only if there is exactly one a-node with the given lemma in both the hypothesis and the reference. Therefore, the MT order might be undefined for some nodes.

In the next step, we compute the *subtree MT order* of each reference a-node R as the average MT order of all a-nodes in the subtree rooted at the a-node R (including the MT order of R itself). Only nodes with a defined MT order are taken into account, so the subtree MT order can be undefined for some nodes.

Finally, we iterate over all a-nodes recursively starting from the bottom. Head a-node $H$ and its dependent a-nodes $D_i$ is reorder if they violate the *sorting order*. If $D_i$ is a root of a subtree, the whole subtree is moved and its internal ordering is kept.

The sorting order of $H$ is defined as its MT order; the sorting order of each dependent node $D_i$ is defined as its subtree MT order. If a sorting order of a node is undefined, it is set to the sorting order of the node that precedes it, thus favouring neighbouring nodes (or subtrees) to be reordered together in case there is no evidence that they should be brought apart from each other. Additionally, each sorting order is added 1/1000th of the original order of the node – in case of a tie, the original ordering of the nodes is preferred to reordering.

### 5.2.5 Synthesis from a-layer to w-layer

The word forms are already generated on the a-layer, so there is little to be done. Superfluous tokens are deleted (e.g. duplicated commas), prepositions are vocalized, the sentence beginning is capitalized, and the tokens are concatenated (a set of rules is used to decide which tokens should be space-delimited and which should not).

The example sentence (from Figure 8) results in the following sentence: *Internet vyvolal boom těchto spekulací.* ("The Internet has caused a boom of these speculations."), which has the same meaning as the original reference sentence, is grammatically correst and most importantly is much more similar in wording to the hypothesis.

### 5.2.6 Results

We evaluate the new paraphrased references with three different metrics – BLEU, Meteor and the Meteor metric without the paraphrase support (based on Section 4.5 and the fact that it seem redundant to use paraphrases on already paraphrased sentences).

The results are presented in Table 6 as a Pearson correlation of a metric with human judgment. Contrary to Moses results (Table 5), paraphrasing using Treex clearly helps to reflect the human perception better.

However, the results are worse than the Sim-

|  | **WMT12** |  |  | **WMT13** |  |  |
|---|---|---|---|---|---|---|
| references | original | paraphrased | reordered | original | paraphrased | reordered |
| BLEU | 0.751 | 0.783 | 0.804 | 0.834 | 0.850 | 0.878 |
| Meteor | 0.833 | 0.864 | 0.870 | 0.817 | 0.871 | 0.870 |
| Ex.Meteor | 0.861 | 0.900 | **0.904** | 0.848 | **0.893** | **0.893** |

Table 6: Pearson correlation of a metric and human judgment on original references, paraphrased references and paraphrased reordered references. Ex.Meteor represents Meteor metric with exact match only (i.e. no paraphrase support).

ple substitution method (Section 4), even though they essentially perform similar task – one-word substitution. One reason of the different performance is smaller number of substitutions, only 1.39 (WMT12) / 1.12 (WMT13) word per sentence. We still inquire the reason of this drop.

The reordering clearly helps when we evaluate via the BLEU metric, which punishes any word order changes to the reference sentence. Meteor is more tolerant to word order changes and the reordering has practically no effect on his scores.

However, manual examination showed that our constraints are not strong enough to prevent creating non grammatical sentences. The algorithm tend to copy the word order of the hypothesis, even if it is not correct. A lot of errors was caused by changes of a word order of punctuation.

## 6 Future Work

Our Treex model as described hardly employs all possibilities of Treex - we only do simple one-word substitutions. In our future work, we plan to extend the paraphrasing pipeline for more complex paraphrases including syntactical paraphrases, multiword phrases, light verbs construction, diatheses, deleting unnecessary words, etc.

We plan to revise the word ordering scheme and add rule-based constrains to stop ungrammatical constructions. Furthermore, we would like to learn automatically possible word order changes from Deprefset (Bojar et al., 2013b), which contains an excessive number of manually created reference translations for 50 Czech sentences.

We also plan to change only parts of sentences that are dependent on paraphrased words, thus keeping the rest of the sentence correct and creating more conservative reference sentences and thus avoiding inaccuracies during synthesis.

We perform our experiment using Treex on Czech language, but the procedure is generally language independent, as long as there is analysis and synthesis support for particular language in Treex. Currently there is full support for Czech, English, Portuguese and Dutch, but there is ongoing work on many more languages within the QTLeap[11] project.

One of our problems is the noise in the paraphrase tables. We plan an automatic filtering and including recently released paraphrase database PPDB (Ganitkevitch and Callison-Burch, 2014). We also intend an experiment of paraphrasing without paraphrase tables based on `word2vec` (Mikolov et al., 2013) similarity; or `word2vec` improved by paraphrase tables (Faruqui et al., 2014).

## References

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Felisa Verdejo. 2009. The Contribution of Linguistic Features to Automatic Machine Translation Evaluation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 306–314.

Colin Bannard and Chris Callison-Burch. 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.

Petra Barančíková, Rudolf Rosa, and Aleš Tamchyna. 2014. Improving Evaluation of English-Czech MT through Paraphrasing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014).*

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague Dependency Treebank 3.0.

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

Ondřej Bojar, Kamil Kos, and David Mareček. 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.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar F. Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 1–11, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013a. 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.

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. 2013b. Scratching the Surface of Possible Translations. In *Text, Speech and Dialogue: 16th International Conference, TSD 2013. Proceedings*, pages 465–474, Berlin / Heidelberg. Springer Verlag.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Ondřej Bojar. 2012. *Čeština a strojový překlad (Czech Language and Machine Translation)*, volume 11 of *Studies in Computational and Theoretical Linguistics*. ÚFAL, Praha, Czech Republic.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006a. Improved statistical machine translation using paraphrases. 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 17–24.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006b. Re-evaluating the Role of BLEU in Machine Translation Research. *Proceedings of EACL*, 2006:249–256.

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

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. In *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR*.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Helge Dyvik. 1998. Translations as Semantic Mirrors: from Parallel Corpus to Wordnet. In *Proceedings of the Workshop Multilinguality in the lexicon II at the 13th biennial European Conference on Artificial Intelligence (ECAI'98)*, pages 24–44, Brighton, UK.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. 2014. Retrofitting Word Vectors to Semantic Lexicons. *CoRR*, abs/1411.4166.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The Multilingual Paraphrase Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, may.

David Kauchak and Regina Barzilay. 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.

Philipp Koehn, 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. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michal Křen, Tomáš Bartoň, Václav Cvrček, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Renata

Novotná, Vladimír Petkevič, Pavel Procházka, Věra Schmiedtová, and Hana Skoumalová. 2010. SYN2010: balanced corpus of written czech.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2011. Approximating a Deep-syntactic Metric for MT Evaluation and Tuning. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 92–98, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 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.

Nitin Madnani. 2010. *The Circle of Meaning: From Translation to Paraphrasing and Back*. Ph.D. thesis, University of Maryland, College Park.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-derived Paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 381–390, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective Dependency Parsing Using Spanning Tree Algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 523–530.

Dennis N Mehay and Michael White. 2012. Shallow and Deep Paraphrasing for Improved Machine Translation Parameter Optimization. *The AMTA 2012 Workshop on Monolingual Machine Translation, MONOMT*.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

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

Preslav Nakov. 2008. Improved Statistical Machine Translation using Monolingual Paraphrases. *ECAI 2008: 18th European Conference on Artificial Intelligence, July 21-25, 2008, Patras, Greece: Including Prestigious Applications of Intelligent Systems (PAIS 2008): Proceedings*, 178:338.

Václav Novák and Zdeněk Žabokrtský. 2007. Feature Engineering in Maximum Spanning Tree Dependency Parser. In Václav Matousek and Pavel Mautner, editors, *TSD*, Lecture Notes in Computer Science, pages 92–98. Springer.

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

Franz Josef Och, Christoph Tillmann, Hermann Ney, and Lehrstuhl Fiir Informatik. 1999. Improved Alignment Models for Statistical Machine Translation. In *University of Maryland, College Park, MD*, pages 20–28.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Labelled Dependencies in Machine Translation Evaluation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sebastian Padó, Daniel Cer, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Measuring Machine Translation Quality as Semantic Equivalence: a Metric Based on Entailment Features. *Machine Translation*, 23(2-3):181–193, September.

Karel Pala, Tomáš Čapek, Barbora Zajíčková, Dita Bartůšková, Kateřina Kulková, Petra Hoffmannová, Eduard Bejček, Pavel Straňák, and Jan Hajič. 2011. Czech WordNet 1.9 PDT.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*, IceTAL'10, pages 293–304, Berlin, Heidelberg. Springer-Verlag.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 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.

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Number v. 6 in Generativní popis jazyka a česká deklinace. Academia.

Matthew G. Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3):117–127, September.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 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.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. pages 901–904.

Jana Straková, Milan Straka, and Jan Hajič. 2014. 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. Association for Computational Linguistics.

Stephanie Strassel, Christopher Cieri, Andrew Cole, Denise Dipersio, Mark Liberman, Mohamed Maamouri, and Kazuaki Maeda. 2006. Integrated linguistic resources for language exploitation technologies. In *In Proceedings of LREC*.

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. TectoMT: Highly Modular MT System with Tectogrammatics Used As Transfer Layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 167–170.

Liang Zhou, Chin yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In *In Proceedings of EMNLP*.