PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 101 APRIL 2014 71–96

Czech Machine Translation in the project CzechMATE

Ondřej Bojar, Daniel Zeman

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

Abstract

We present various achievements in statistical machine translation from English, German, Spanish and French into Czech. We discuss specific properties of the individual source languages and describe techniques that exploit these properties and address language-specific errors. Besides the translation proper, we also present our contribution to error analysis.

1. Introduction

This overview article summarizes recent advances in machine translation involving Czech, as achieved within the project CzechMATE. The overall state of the art in machine translation has recently made a leap in quality, esp. with the introduction of phrase-based methods (Koehn, 2004; Koehn et al., 2007) and availability of very large corpora (Bojar et al., 2012). The output may still suffer many errors, including serious ones such as completely reversed meaning, but it is nevertheless rather difficult to improve it fully automatically. The main reason is that the spectrum of outstanding errors is diverse and the errors are not easy to fix without negatively affecting the rest of the sentence. Aware of this challenging situation, we experimented at multiple fronts, searching for the "lower-hanging fruit".

The article has two main parts: Section 2 is concerned with methods of machine translation evaluation, covering techniques that fully rely on human judgement, techniques fully automatic as well techniques and tools mixing the two. Section 3 describes a wide range of our experiments with statistical machine translation into Czech, be it small specific ideas such as handling of named entities or breaking words into morphemes, or a complex combination of three big components into the current best English-to-Czech MT system.

© 2014 PBML. All rights reserved.
Corresponding author: zeman@ufal.mff.cuni.cz
Cite as: Ondřej Bojar, Daniel Zeman. Czech Machine Translation in the project CzechMATE. The Prague Bulletin of Mathematical Linguistics No. 101, 2014, pp. 71–96. doi: 10.2478/pralin-2014-0005.

2. Novel Methods of Machine Translation Evaluation

Measuring progress is a critical component of scientific work but evaluating machine translation is unfortunately a rather peculiar task.

When comparing two possible MT outputs relative to each other, be it two distinct MT systems or an older vs. a newer version of a single system as a progress check, the hypotheses are often incomparable. One sentence can have an error in the beginning, the other at the end. One can be more or less correct but disfluent, the other can be perfectly fluent but reverse the meaning.

When assessing the quality of a single hypothesis on an "absolute" scale, we face subjectivity of human judgments: if the output of the MT system is not error-free right away, it is usually not clear which part is wrong because an input sentence has many possible translations, see also Section 2.5. Also, some errors are more serious than others, e.g. some distortion of the meaning vs. a clear and non-confusing typo or a typographical error. Moreover, each annotator gets quickly accustomed to the errors of the system and it is increasigly hard for him to notice them.

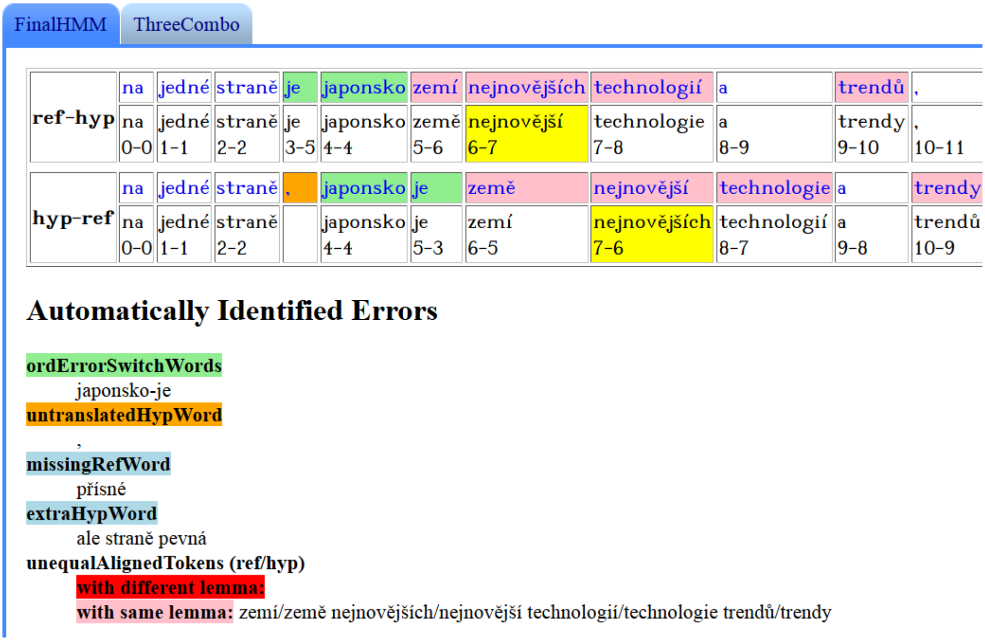
The field of MT evaluation is thus actively evolving and no completely satisfactory solution has been found so far. In our project, we contributed to both manual (Sections 2.1, 2.2 and 2.3) and automatic evaluation methods (Sections 2.4 and 2.5).

2.1. Tools for Manual MT Output Inspection and Analysis

We developed an interactive tool for analysis of MT errors, called Addicter (Automatic Detection and DIsplay of Common Translation ERrors) (Berka et al., 2013). It can do the following:

- Find erroneous tokens and classify the errors in a way similar to Vilar's taxonomy (Vilar et al., 2006). This component for automatic error detection and classification is further discussed in Section 2.4.
- Browse the test data, sentence by sentence, and show aligned source sentence, reference translation and system hypothesis (Figure 1).
- Browse aligned training corpus and look for example words in context.
- Show lines of the phrase table that contain a given word.
- Summarize alignments of a given word. This feature can also serve as a primitive corpus-based dictionary.
- Search and group words sharing the same lemma. That way, morphological errors can be highlighted.

The test data browser facilitates examination of system-generated hypothesis and its comparison to the reference translation(s). On the other hand, the search engine that operates on training corpus and phrase table can reveal whether an out-of-vocabulary word really never occurred in training data, or it got filtered out during subsequent processing.



ref-hyp	na	jedné	straně	je	japonsko	zemí	nejnovějších	technologií	a	trendů	,
	na 0-0	jedné 1-1	straně 2-2	je 3-5	japonsko 4-4	země 5-6	nejnovější 6-7	technologie 7-8	a 8-9	trendy 9-10	, 10-11
hyp-ref	na	jedné	straně	,	japonsko	je	země	nejnovější	technologie	a	trendy
	na 0-0	jedné 1-1	straně 2-2		japonsko 4-4	je 5-3	zemí 6-5	nejnovějších 7-6	technologií 8-7	a 9-8	trendů 10-9

Automatically Identified Errors

ordErrorSwitchWords

japonsko-je

untranslatedHypWord

,

missingRefWord

přísné

extraHypWord

ale straně pevná

unequalAlignedTokens (ref/hyp)

with different lemma:

with same lemma: zemí/země nejnovějších/nejnovější technologií/technologie trendů/trendy

Figure 1. Screenshot of the Test Data Browser in Addicter. Different types of errors are highlighted using different colors. Monolingual word alignment between reference translation and system hypothesis is indicated using numerical indexes of words, and also highlighted when the mouse pointer is over a word. The English source of the sentence in the example is “On the one hand, Japan is the land of the latest technologies and trends, but on the other hand it is strict, disciplined and traditional.”

2.2. Evaluating Predicate-Argument Structure

An established manual evaluation method asks annotators to RANK up to five different MT outputs for a given input sentence. The method is not sufficiently reliable (see also Section 3.7 below) but it has nevertheless been in use for a long series of WMT evaluation campaigns, see Callison-Burch et al. (2007) through Bojar et al. (2013a). The assignment for the ranking method is very simple: “You are shown a source sentence followed by several candidate translations. Your task is to rank the translations from best to worst (ties are allowed).” This simplicity is flexible with respect to overall system quality and allows to include untrained annotators in the evaluation but comes at a cost: people may focus on different aspects of the outputs, and even a single annotator may use inconsistent “quality scales” across sentences.

Reference	Oblečky	musíme	vystříhat	z časopisů
Gloss	clothes	we-must	cut	from magazines
Roles	Experiencer	Modal	Action	Locative
Meaning	We must cut the clothes (assuming paper toys) from magazines			
Hypothesis	Musíme	vyříznout	oblečení z časopisů	
Gloss	We-must	cut	clothes from magazines	
Roles	Modal	Action	Experiencer	

Figure 2. HMEANT: labelling semantic roles of phrases in the reference translation and the hypothesis. The example illustrates a problem of PP-attachment mismatch. One of the annotations treats the phrase "z časopisů" correctly as a separate frame filler, labelling it as Locative, the other annotation groups it together with the noun "oblečení" into an Experiencer. The subsequent alignment of slot fillers is not possible since two slots in the reference (Experiencer and Locative) correspond to one slot in the hypothesis (Experiencer).

Lo and Wu (2011) propose HMEANT, a manual method that evaluates MT output based on the core predicate-argument structure of the sentence. HMEANT checks, whether key elements of clause structure were preserved: "who did what to whom, how where and why". HMEANT was designed with future full automation in mind and its principles are very much in line with the structuralist linguistic theory (Sgall et al., 1986; Panevová, 1994; Lopatková et al., 2008) and processing tools (Popel and Žabokrtský, 2010) as developed at our institute. We thus decided to apply HMEANT to Czech, the first language other than English where HMEANT has been tested so far.

HMEANT consist of two phases:

Semantic role labelling where both the reference and the MT output receive an explicit annotation of which word is the predicate and which groups of words correspond to its arguments (or "semantic role fillers").

Alignment where predicates and their arguments in the hypothesis and in the reference get aligned whenever possible.

The final formula of HMEANT then evaluates the f-score of the match between the predicates and their arguments in aligned sentences.

Our experiment, as detailed in Bojar and Wu (2012), confirmed that annotators feel more confident when following the relatively simple HMEANT guidelines compared to the standard hypothesis ranking. On the other hand, we identified a number of issues with the current design of HMEANT and we proposed various changes to the guidelines. One example is depicted in Figure 2: the role labelling, done before alignment, sometimes produces annotations that cannot be aligned unless HMEANT allows for up to many-to-many alignments between arguments.

A subsequent independent experiment by Birch et al. (2013) observed similar problems for English and German and, more importantly, revealed that the inter-annotator agreements on individual labellings (the identification of the predicate, the filler spans as well as their semantic role labels) is relatively good but the alignment is problematic and reaches much lower agreement rates. Also, the discriminatory power of HMEANT is sometimes too weak, since only verbal frames are considered.

In conclusion, HMEANT is a step in a promising direction of MT evaluation but several issues need to be addressed.

2.3. Quiz-Based Evaluation

Most of non-comparative MT evaluation methods focus on whether the translation conveys the general meaning of the original and/or whether it satisfies various linguistic conditions in the target language. Some tests also check the general understandability of the sentence: "Tell us, what the sentence says", validated by a second annotator (Callison-Burch et al., 2009, 2010; Bojar et al., 2013a). In many practical situations, the output of MT can be nevertheless useful even if it contains many errors and omissions. We wanted to try also this much more applied type of evaluation.

Running a genuine task-based evaluation, e.g. bringing an object that matches a machine-translated description, is very expensive. We thus opted for a lighter variant: annotators were given short text snippets translated from native English to Czech by one of four evaluated MT systems. The task was to answer three Czech yes/no questions about the snippet.

Table 1, reproduced from our detailed description of the experiment (Berka et al., 2011), documents that different evaluation methods promote different systems. Specifically, the deep-syntactic system TectoMT in 2010 (see also Section 3.7) seemed to perform relatively poorly in terms of general understandability or n-gram based metrics like BLEU (Papineni et al., 2002) but it supported best our annotators in answering quiz questions.

It is worth noting that quiz-based evaluation scores are generally much higher than the other manual scores (all manual scores range from 0 to 100), which indicates that finding the answer to a specific question in machine-translated text is easier than understanding the whole sentences. Part of this result can be attributed to the fact that the question itself may help a lot in understanding the MT output and also that it may ask for just one or two words of the output. Quiz-based evaluation is thus perhaps insufficiently discerning but it brings some optimism towards the practical utility of machine translation.

2.4. Automatic Error Identification

Automatic error classification in Addicter is based on finding erroneous words in the translation output and assigning a corresponding error class to each of them. The

Metric	Google	CU-Bojar	PC Translator	TectoMT
$\geq$ others (WMT10 official)	**70.4**	65.6	62.1	60.1
$>$ others	49.1	45.0	**49.4**	44.1
General understandability [%]	**55**	40	43	34
Quiz-based evaluation [%]	80.3	75.9	80.0	**81.5**
BLEU	**0.16**	0.15	0.10	0.12
NIST	**5.46**	5.30	4.44	5.10

Table 1. Manual and automatic scores of four MT systems taking part in the WMT10 evaluation campaign. Best results in bold. We report WMT manual evaluations (comparison with other systems and general understandability), our quiz-based evaluation and two automatic scores: BLEU and NIST. Note that the set of considered sentences was not identical across the evaluation methods for various reasons but the actual MT systems were.

user gets an overall picture of the output quality in the given test set and they can start by exploring the most frequent error types first.

The algorithms are described in more detail in (Zeman et al., 2011). The first step is the automatic detection of errors in the hypothesis translation. This is done by comparing tokens in the hypothesis with the reference translation, relying on some word alignment between the two texts.

Previous experience, i.a. (Bojar, 2011), shows that the quality of the alignment is critical, otherwise many errors can be mis-classified (e.g. a pair of "missing" and "extra" errors instead of one error of bad lexical choice). Our tool circumvents current limitations of word alignment by providing several methods and allowing the user to choose which of them works best for the particular language pair, MT system and dataset.

Currently, Addicter internally implements the alignments of WER (Levenshtein, 1966), LCS (Hunt and McIlroy, 1976), a greedy injective alignment, and an injective HMM (Fishel et al., 2011). The user can provide any additional alignment, e.g. GIZA++ (Och and Ney, 2003) alignments, or alignments obtained by linking a reference-to-source alignment with the source-to-hypothesis alignment as reported in a verbose output of the MT system (briefly called "via-source").

Addicter tries to automatically classify errors into categories similar to those of (Vilar et al., 2006), such as: morphological, reordering, missing words, extra words and lexical errors.

The errors can be then summarized into a table showing their counts in the test data. When using GUI, the table is connected to the test data browser, so with just one click, the user can see the list of sentences with the occurrence of the given error type and even look at the sentences one by one in more detail.

	Number of		Correlation
Test Set Origin	Sentences	References	BLEU vs. Manual Rank
WMT11 Official	3003	1	0.69
WMT11 Official	50	1	0.47
Post-edited MT Output	50	1	0.72
BLEU "Standard"	3003	4	0.74
Many Refs	50	500–50k	0.79

Table 2. Performance of BLEU (in terms of correlation with manual ranking of systems) depending on the test size (number of sentences) and the number of references per sentence.

2.5. Very Large Number of Reference Translations

It is well known that n-gram-based automatic evaluation methods notoriously depend on the number of available reference translations. A single reference, although most often used, cannot account for the range of correct possible outputs. For morphologically rich languages like Czech, the situation is worse; see also Bojar et al. (2010). The proper use of BLEU (Papineni et al., 2002) requires 4 references.

Dreyer and Marcu (2012) proposed to manually construct many, if not all, reference translations using a compact representation. In Bojar et al. (2013b), we aim at the same goal with Czech, which required us to completely redesign the framework to accommodate linguistic properties of Czech. The project CzechMATE then paid a few annotators to construct compact representations of many Czech references of just 50 English sentences. We confirm that it is easy to produce dozens or hundreds of thousands reference translations per one input sentence.

Table 2 summarizes our results. The correlation of manual ranking and BLEU with just one reference as provided for WMT evaluation campaigns is not very high: 0.69. If we restrict the testset to just the exact 50 sentences we use, the correlation drops to unacceptable 0.47. Having a reasonably sized test set is thus very important if only one reference is available; however, the situation seems to get much better if the 50 references are constructed by post-editing MT outputs and not translated independently: 0.72.

The proper use of BLEU (4 references over a 3000-set of sentences) performs acceptably: 0.74. We are nevertheless able to improve on this and reach 0.79 if we use many references. As we are adding references, the correlation steadily improves and it saturates around 500 reference translations selected randomly from the tens of thousands of possibilities. Due to the random selection, this 500-reference set is probably more diverse than if asked translators to produce 500 different translations.

We find the approach to MT evaluation which relies on (very) many reference translations and an automatic metric very promising since it mitigates most problems of manual MT evaluation (subjectivity, difficult replicability). Currently, constructing the many references is prohibitively expensive (2 hours per sentence) but we hope techniques to speed this up will emerge.

3. Improving Statistical MT into Czech

For a long time, published research on translation into Czech strongly focused on translation from English. Arguably, translation from different source languages may pose different problems. We picked a few other major European languages (German, Spanish and French) and compared English-Czech translation with translation from these languages. The choice was motivated practically both from the point of view of the user (expected likelihood that a user will need such translation) and of the developers (availability of training data). We are aware that the difference between English and any of these three languages is almost negligible when compared to the difference between English and any non-European language (e.g. Chinese). However, even our restricted experimental environment proves that individual, linguistically driven approach to every language pair is desirable and beneficial.

3.1. Core System

The findings presented in this article are based on experiments with various MT systems that differ from each other in their settings, combination of training data, pre- and postprocessing steps etc. Nevertheless, there is a core technology common to most of these systems. Unless explicitly stated otherwise, we use the Moses SMT decoder (Koehn et al., 2007) with baseline settings, Giza++ to compute word alignments (Och and Ney, 2003), the well-known *grow-diag-final-and* symmetrization heuristic (Koehn et al., 2003), the SRILM language modeling toolkit (Stolcke, 2002) and the minimum error-rate training algorithm, MERT (Och, 2003).

3.2. Data

We need two sorts of corpora for statistical machine translation: parallel data for translation models and monolingual data for target language models. The CzEng 1.0 parallel corpus (Bojar et al., 2012) provides a decent amount of English-Czech parallel data. It is less easy to obtain training data for German-to-Czech, Spanish-to-Czech and French-to-Czech translation.

We used the data provided for the annual WMT translation task (Callison-Burch et al., 2012)[1]: the Europarl corpus and the News Commentary parallel corpus. Both of the corpora contain text in each of the five languages we are interested in (Czech, English, German, Spanish and French). However, not all segments are available in all languages and the corpora are supplied as four Something-English pairs. Fortunately

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

there is a significant overlap across the pairs, so we were able to combine them. For instance, to create a Czech-German parallel corpus, we identified the intersection of the English sides of Czech-English and English-German corpora, respectively; then we combined the corresponding Czech and German sentences.

Table 3 shows the sizes of the corpora.

Corpus	SentPairs	Tokens lng1	Tokens lng2
en-cs	782,756	20,964,639	17,997,673
de-cs	652,193	17,422,620	15,383,601
es-cs	692,118	20,189,811	16,324,910
fr-cs	686,300	22,220,780	16,190,365
de-en	2,079,049	55,143,719	57,741,141
es-en	2,123,036	61,784,972	59,217,471
fr-en	2,144,820	69,568,241	59,939,548
CzEng en-cs	14,833,358	231,463,445	200,724,410

Table 3. Number of sentence pairs and tokens for every language pair in the Europarl, News Commentary and CzEng corpora. Every line corresponds to one language pair from combined Europarl and News Commentary, except for the line marked as CzEng. Languages are identified by their ISO 639 codes: cs = Czech, de = German, en = English, es = Spanish, fr = French.

Furthermore, there are test sets from six years of WMT shared tasks, about 3,000 sentences each. These are multi-parallel, available in all five languages. The contents are short news stories originally written in one of the languages (balanced) and human-translated to the others. We use these sets as development and test data; we hired translators from German to Czech to obtain four different Czech reference translations of the WMT 2011 set.[2]

The Czech side of the parallel corpora, especially CzEng, provides a decently sized monolingual corpus for training the target language model. Its size is still suboptimal for a morphologically rich and free-word-order language such as Czech. The situation would be much better for translation out of Czech, with Gigaword corpora (Parker et al., 2011) available for English, French and Spanish; unfortunately, there is no Czech Gigaword corpus yet. We use the Czech Crawled News corpus instead (also provided by WMT). It consists of 460 million tokens in 27.5 million segments (sentences).

All parallel and monolingual data underwent the same preprocessing. They were tokenized the same way, a few special characters were normalized or cleaned, and a set of language-dependent heuristics was applied in an attempt to restore and nor-

[2] These additional reference translations also served in our experiment described in Section 2.5 and they are freely available to other researchers, see `http://hdl.handle.net/11858/00-097C-0000-0008-D259-7`.

malize the opening/closing quotation marks (i.e. "quoted" → “quoted”). First, we hope that paired quotation marks could occasionally work as brackets and better denote parallel phrases for Moses; second, if Moses learns to produce directed quotation marks, subsequent detokenization will be easier.

Our heuristics applied to 1.84 % of Spanish sentences, 2.47 % Czech, 2.77 % German, 4.33 % English and 16.9 % French (measured on Europarl data). See Zeman (2012) for details.

We tag and lemmatize all texts. Lemmas are used to compute word alignment, and also to apply “supervised truecasing” (upper- or lowercase of the first letter of a word form is derived from its lemma). Without supervised truecasing, the models could not correctly utilize sentence-initial words, which are always uppercased. In Bojar et al. (2013c), we empirically confirm that this supervised truecasing indeed performs best for English-to-Czech translation. Morphological tags are used for delexicalized language modeling, to assess fluency of a morphologically rich language. We use the Featurama tagger for Czech and English lemmatization and TreeTagger for German, Spanish and French lemmatization. All these tools are embedded in the Treex analysis framework (Žabokrtský et al., 2008).

3.3. English as Pivot Language

Table 3 demonstrates that there are significantly more data for German / Spanish / French-English and for English-Czech translation, than directly for German / Spanish / French-Czech. Since more data typically means better models, one should ask whether we would not do better if we translated the source text first to English, then from English to Czech. There are also some obvious drawbacks of such an approach: since MT systems typically make errors, applying a system twice in a row could accumulate more errors. Every language has its own specific properties, be it fixed word order, morphological features or parts of speech that do not occur in other languages. Making the output consistent with these specifics is one of the biggest challenges that every MT system faces. Therefore, having first to accommodate constraints of an intermediate language could make the task unnecessarily difficult.

We ran several experiments to see which of the two approaches is better. Results are shown in Table 4. In terms of BLEU score, the results are not equally conclusive over all language pairs. The intermediate level clearly hurts German-Czech translation, to a lesser extent it also damages Spanish-Czech results. French-Czech seems to be (as far as BLEU can tell) the most difficult one among the investigated language pairs, and intermediate English does not change it (the fr-en-cs BLEU score is even higher than fr-cs but the difference is not statistically significant).

The table also shows human evaluation of the experiments, which provides a different picture. It suggests that both German and French are better translated via English (exploiting the big models). Spanish appears to be the least sensitive to the choice

src fr	les diabétiques ne seront plus tenus de contrôler leur taux de sucre .
fr-cs	na diabétiques nebudou povinny dohlížet na jejich cukru .
fr-en-cs	diabetiků nejsou povinny monitorovat jejich podíl cukru .
en-cs	diabetiků už nebudou muset kontrolovat hladinu cukru v krvi .
ref cs	diabetici si již nemusí hlídat hladinu cukru .
ref en	diabetics no longer need to control their blood sugar .

Figure 3. French to Czech direct or via English. The word diabétiques is OOV in the direct fr-cs model. The larger models with pivot English managed to find a Czech equivalent, even though they failed to pick the correct form (genitive diabetiků instead of nominative diabetici).

src de	Sie forderten weiterhin die Bildung von Gewerkschaften in der fast - Food - Branche .
de-cs	žádali i nadále vzdělání odbory v rychlé občerstvení - odvětví .
de-en-cs	žádali , aby mohli pokračovat v zakládání odborů ve fast - food industry .
en-cs	také požadovala vznik odborů ve fast food industry .
ref cs	také požadovali , aby ve sféře rychlého občerstvení byly založeny odborové organizace .
ref en	they also demanded the creation of unions in the fast food industry .

Figure 4. German to Czech direct or via English. Pivot English favored “zakládání odborů” (creation of unions) over “vzdělání odborů” (education of unions). Both creation and education are possible translations of German “Bildung”. The direct model was more successful in translating “the fast food industry” but the overall fluency and understandability of the sentence is much better in the pivoted translation.

here; indeed, Spanish was the language where we did not identify many language-specific problems affecting translation into Czech.

Human inspection of the outputs revealed that English often helped to select better target words, or even cover source words that would be OOV (out-of-vocabulary) in the direct model. See the examples in Figure 3 and Figure 4.

3.4. Properties of Individual Source Languages

In this section we summarize linguistic characteristics that are specific to each investigated language pair and that influence the quality of MT output. For each specific phenomenon, we describe a linguistic transformation that, if applied both to the source-language part of the training data and of the test input, makes learning of the translation model easier.

	BLEU		Humans Prefer		
Pair	Direct	Via en	Direct	Neither	Via en
en-cs	0.1786				
de-cs	**0.1532**	0.1334	21	50	**29**
es-cs	**0.1614**	0.1570	**25**	57	18
fr-cs	0.1441	**0.1466**	25	39	**36**

Table 4. BLEU-score and manual evaluation of translation from various source languages to Czech. The figures in the first column evaluate direct models between the source language and Czech, trained on small data. The second column is via English, where much larger data is available for both steps (source to English and English to Czech). The last column shows the percentage of cases where human judgement scored the direct translation better, equally good (bad), or worse than the translation via English.

Czech is the target language in all four translation pairs; Czech has rich morphology (both inflectional and derivational) and relatively free word order. In fact, the predicate-argument structure, often encoded by fixed word order in English, is usually captured by inflection (especially the system of 7 grammatical cases) in Czech. The non-English source languages have freer word order and more morphology than English but still their morphology is much simpler than in Czech. Generating correct inflectional affixes is thus one of the main challenges of translation from any of these languages into Czech. Furthermore, the multitude of possible Czech word forms (significantly higher than in English) makes the data sparseness problem really severe.

The grammar of Czech requires two main layers of grammatical agreement: subject agrees with verb in person, number and gender, and adjective agrees with its governing noun in gender, number and case. (Furthermore, the choice of case is controlled by valency of verbs and prepositions.) Language models are typically too weak to enforce the agreement reliably. One of the most common translation errors is a wrong morphological form of otherwise correctly picked lemma.

A less pronounced difference between Czech and all the source languages is that Czech normally does not mark definiteness: there are no definite or indefinite articles. It is easy for the MT system to drop the articles; however, learning phrases with two different articles (or without any article) unnecessarily disintegrates statistics and makes the phrase table larger. Experiments show that dropping articles during preprocessing of the training data simplifies the models without decreasing the BLEU score. (Unlike in English, the German, Spanish and French articles also distinguish gender. It is not of much use for translation though, as the grammatical gender in the target language may differ.)

The following subsections present ideas how to adapt models to the individual source languages; these ideas have yet to be verified in experiments.

3.4.1. English to Czech

Czech is a pro-drop language, i.e. it is not required to supply a personal pronoun whenever there is no better subject in the sentence. However, Czech finite verbs are marked for person, which is much less visible in English. We can design a preprocessing step that will make sure that there is always a personal pronoun next to the English verb—even if there is a noun phrase functioning as subject. It will help the translation model to learn the correct Czech verb forms.

One area where the English language system is much more complex than the Czech one, is tenses. Czech has only three tenses (past, present and future). No perfect tenses (there are special perfective and imperfective verbs) and no progressive tenses. The periphrastic verb forms in English are a common source of translation errors. For instance, the auxiliary "is" (as in "he is doing") is sometimes translated to Czech, although it should not be used there. We thus need a preprocessing step that identifies the tense of the English verb and, if necessary, maps it to simple past, present or future. This way auxiliaries will be seen only with future and the trainer will find it easier to learn translation of content verbs.

In our combined system (Section 3.7), the complex English tenses are specifically handled by the deep-syntactic system TectoMT.

3.4.2. Spanish to Czech, French to Czech

Out of our language pool, Spanish and French possess the least grammatical peculiarities (but see the notes on negation in Section 3.4.4). Their word order is mostly compatible with preferred Czech word order, with one important exception: Adjectival modifiers usually follow the noun, in Czech they precede it.

3.4.3. German to Czech

German is genetically related to English (both belong to the Germanic group) and it has long history of close neighborhood and influencing of Czech. Nevertheless, it is distinctively different from both.

The uppercase / lowercase distinction is more important in German than in other languages because all German nouns (not just named entities) are capitalized.

Long German compound words are notorious for increasing out-of-vocabulary rate, which has led many researchers to devising unsupervised compound-splitting techniques. For instance, the word "Geschichtenerzähler" (storyteller) is OOV in our data; if we split it to "Geschichten" (stories) and "Erzähler" (teller), neither part will be OOV (see Section 3.5 for our approach to compound splitting).

German word order is not as fixed as English but there are strict rules about placement of verbs. Dependent verbal forms including participles are placed at the end of the clause, and the resulting long-distance dependencies often have deadly effect on meaning preservation, as demonstrated in Figure 5.

src de	französische Truppen haben ihren Verantwortungsbereich verlassen
lit. en	French troops have their responsibility-area left
ref en	French troops have left their area of responsibility
ref cs	francouzské jednotky opustily svou oblast odpovědnosti
de-cs	francouzské jednotky mají své povinnosti
lit.	French troops have their responsibility

Figure 5. Source German sentence uses the present perfect tense, "haben verlassen" (have left); however, the participle is placed far away from the finite form of "haben". The finite verb serves as auxiliary in this sentence but it could act as content verb elsewhere. Our model overlooked the participle at the end and took "haben" for content-bearing verb. As a result, "to leave responsibility area" was misinterpreted as "to have responsibility".

src es	en estos nuevos tiempos no es cómodo trabajar
lit. en	in these new times not is comfortable to-work
ref en	in the new times, it is uncomfortable to work
es-cs	v této nové době je příjemné pracovat
lit. en	in this new time is pleasant to-work
ref cs	v nových časech je práce nepříjemná

Figure 6. Negation lost in translation. Czech negation is typically marked by a morpheme bound to the verb, not by a separate particle. In this case, the words "no es" should be translated as one word "není", not "je" (underlined); however, since there are various other means to express the negative meaning (one of them instantiated in the Czech reference here), the model learned that the negative particle "no" often remains unaligned. It is a dangerous observation and leads to dropping of negation quite frequently.

The problem can be solved by moving participles back to the auxiliaries during preprocessing of the data, see i.a. Collins et al. (2005). Similarly, one may also want to move separable verb prefixes closer to the corresponding verb stems.

3.4.4. Negation

The various linguistic devices for expressing negation pose a separate set of problems. It is possible to generate a perfectly fluent sentence with 95 % words translated correctly, yet with the overall meaning totally reversed (e.g. Figure 6) – this is also one of the reasons of low reliability of many automatic MT evaluation methods.

Czech negation is typically marked using the prefix "ne-" on verbs or adjectives (example: "Student **ne**přišel.") In English, the auxiliary verb "to do" is usually needed

	BLEU		Humans Prefer		
Pair	Words	Morphs	Words	Neither	Morphs
en-cs	**0.1632**	0.1425	29	**42**	29
de-cs	**0.1532**	0.1272	24	33	**43**
es-cs	**0.1614**	0.1344	**31**	45	24
fr-cs	**0.1441**	0.1186	**31**	40	29

Table 5. BLEU-score and manual evaluation of translation using a word-based model (default) and a morpheme-based model. All models were trained on combined News Commentary and Europarl (no CzEng), additional language model was trained on the Czech Crawled News corpus. Morphemes were recombined to words before evaluation of the morpheme-based models. The three columns to the right show the percentage of cases where human judgment scored the word-based translation better, equally good (bad), or worse than the morpheme-based translation.

("The student **did not** come.") But adjectives behave the same way as in Czech: the prefix **un-** is a bound morpheme. In German and Spanish there is a negative particle but no auxiliary verb is needed ("Der Student kam **nicht.**" "El estudiante **no** llegó.") French is different in that it has two negative particles ("L'étudiant **n'**est **pas** venu.")

As with other peculiarities, the training situation can be improved using a cheap trick: separate the Czech negative prefixes so that they are learned as separate words. It will reduce the number of occurrences of unaligned negative particles on the source side. In English, the auxiliary "does / do / did" could be removed (see also progressive tenses in Section 3.4.1).

There is still one problem open, though: it is not rare in Czech to see negation marked on several words in the same clause ("**nikdy ne**přišel **žádný** student" = lit. "never not-came no student"). Such a Czech sentence is difficult to generate because in our source languages the negation is typically marked only once ("**no** student ever came", "the student **never** came", "**ningún** estudiante llegó" etc.) Sentences such as "***Žádný** student přišel." are not grammatically correct Czech. One could attempt to recognize all "negatable" words in the preprocessing phase and negate their source-language version. It is difficult though to identify the exact set of such words and to produce their negated form reliably.

3.5. Morphemic Segmentation of Words

There are many long compounds in German. Many are OOV (unknown from training data, appear in test data) and the set of possible compounds is in theory infinite. It is therefore desirable to split compounds to individual stems during the preprocessing phase. We decided to go a step further and to approach morphological forms in all the languages the same way. We first segment all training words into morphemes,

src es	aquí vislumbramos las premisas de una teocracia » .
ref en	this suggests the beginnings of a theocracy . ”
es-w-cs	zde vislumbramos předpoklady k teokracii . “
es-m-cs	zde vidíme výchozí teokracii “ .
ref cs	lze v tom spatřovat začátky teokracie “ .

Figure 7. Spanish “vislumbramos” (“we see/sense”) is unknown to the word-based model but the morpheme-based model succeeds in decomposing and translating it. It has been segmented as “vislumbr/STM +a/SUF +mos/SUF”. The phrase table contains 700 entries with the stem “vislumbr” but none of them is the 1st person plural “vislumbramos”.

src de	in der römischen Zeit war Caesarea die wichtigste Stadt Judäas
ref en	during the Roman period , Caesarea was the main city of Judea
de-w-cs	v římské době bylo Caesarea hlavní město Judäas
de-m-cs	v římské době bylo Caesarea hlavní město Judäasské
ref cs	v římských dobách byla Caesarea hlavním městem Judeje

Figure 8. The morpheme-based model constructed adjective from Judea. This experiment did not use the model for named entities described later in this article, so Judäa remained in its German spelling but the Czech adjectival suffix “ské” was attached to it.

then learn a translation model from the parallel sequences of morphemes. We hope to model morphological behavior of the other languages too. For example, locative expressions will often be translated to Czech using the locative case. While the system is likely to observe all possible locative suffixes in the training data, it is much less likely to encounter all words in all cases (including locative). But it may be able to combine a known stem with a known locative suffix and create a valid word form, which has not itself occurred in the training data.

We use the freely available tool Morfessor (Creutz and Lagus, 2007) to segment all corpora into morphemes. In addition to segmentation, Morfessor also classifies the morphemes as prefixes, stems and suffixes, respectively. To give an example, the German phrase “aus dem Strafgesetzbuch zu entfernen” (“to remove from the Criminal Code”) is broken down into “aus/STM dem/STM straf/PRE+ gesetz/STM + buch/STM zu/STM ent/PRE+ fernen/STM”. The phrases that Moses learns are sequences of tagged morphemes, thus the Moses decoder also generates similar sequences. We take the generated morphemes, remove their tags and join them on plus signs to get full words again.

src de	das heißt , dass Ibrahimovic leistungsstark ist und viele Tore schießt .
ref en	this means Ibrahimovic is very successful and scores a lot of goals .
de-w-cs	to znamená , že Ibrahimovic výkonná je mnoho a střílí góly .
de-m-cs	to znamená , že ibrahim , ovic silný a mnoho gól střílí .
ref cs	to znamená , že Ibrahimovič je velmi výkonný a že dává hodně gólů .

Figure 9. Sometimes the generated morphemes do not allow for correct rejoining of the word.

src en	diabetics no longer need to control their blood sugar .
en-w-cs	diabetiky už třeba kontrolovat jejich krevního cukru .
en-m-cs	diabetiky už nepotřebují kontrolovat jejich krevního cukru .
ref cs	diabetici si již nemusí hlídat hladinu cukru .

Figure 10. Negation is preserved by the morpheme-based model while the word-based model destroys it.

Table 5 evaluates experiments with morpheme-based translation models. The BLEU scores are disappointing: segmentation to morphemes consistently and significantly hurts the results across all source languages. However, part of the decrease could probably be attributed to the known imperfections of the BLEU metric, as it does not always correlate with human judgment. The difference between words and morphemes is much less pronounced under human evaluation, and in German-to-Czech translation morphemes are even preferred over words. We have not tested the partial model where only the source-language text is segmented.

Figures 7 through 10 portray more details about the pros and cons of the two approaches.

3.6. Named Entities

There is no single and clear-cut definition of what consists a named entity (a name) in a text. It is nevertheless obvious that a high-quality MT system has to address names somehow specifically. Pure cooccurrence statistics in parallel data may be rather deceiving, see the following output of Google Translate:

(1a) Source: Doktor Novák také přišel.
(1b) Google: Dr. Smith also attended.
(1c) Reference: Dr. Novák also came.

Novák, being the most frequent Czech name, is often translated as *Smith* in literary texts. In most other types of text, names should be preserved.

(2a) Source: Sejdeme se v Plzni.
(2b) Google: Meet me in London.
(2c) Reference: Let’s meet in Pilsen.

This second example is just obscure and reveals the level of noise in Google’s parallel data. The actual output differs if we omit the full stop in the source sentence. (Outputs show as of January 2014.)

For our experiments, as detailed in Hálek et al. (2011), we defined a named entity as *a word or group of words which, when left untranslated, are a valid translation anyway*. Some named entities, esp. geographical names and names of institutions, have translations. Depending on the salience of the item, the translation can be preferred very strongly (*Paris–Paříž*), less so (*Trier–Trevír*) or it can become even confusing to people who do not know the specifics (*Görlitz–Zhořelec*), in which case the system should probably produce both variants of the name. For most named entities, there is no counterpart in the other language.

Even if the named entity is not translated, morphologically rich languages like Czech require the entity to be adapted for the context of the sentence (*We met in Trier–Sešli jsme se v Trieru*, or to be introduced with a common noun describing the entity type (*I bought this in IKEA.–Koupil jsem to v obchodním domě IKEA.*, lit. “in the shopping mall IKEA”). While the latter option is also very interesting from the technical point of view (automatically adding the descriptive noun phrase), we attempted to improve the translation and declension of named entities.

We used Wikipedia as the natural source for named entities. This is the outline of our procedure:

1. Search for named entities in the source text. We preferred our simple recognizer based on letter case due to its higher recall over an established recognizer of a higher precision.
2. Confirm entities in Wikipedia. The potential entities need to be present in English Wikipedia and the category of their article has to fall among those that we consider named entities, e.g. Places, People, Organizations.
3. Extract base translation from Wikipedia. We simply follow the cross-language link from the Wikipedia article to find the lexical form of the named entity.
4. Extract variants of the translation. We use the Czech page and collect all phrases with stems identical to the stems of the lexical variant of the named entity. Each of the variant gets a score based on its frequency in the page.
5. Extend the list of “translation options” for the source phrase with all the extracted variants of the translation.

The experiment revealed that automatic MT evaluation disprefers our changes but manual evaluation suggests that both translating entities using our procedure (preferred in 34% cases) as well as identifying them and preserving untranslated (37%) is preferred over the baseline (preferred in 29% of cases).

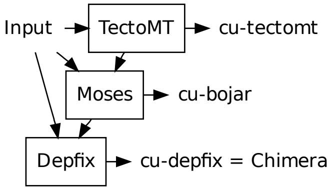


Figure 11. Chimera, a combination of three approaches to English-to-Czech MT.

3.7. Best of All Worlds: System Combination

It has already been established that "system combination" (Matusov et al., 2008), i.e. MT output constructed by combining outputs of several primary MT systems, outperforms the individual systems (Callison-Burch et al., 2009, 2010, 2011).

Our system Chimera described in Bojar et al. (2013c) is constructed differently from the standard system combination techniques. Instead of collecting complete outputs of several systems and selecting words that the majority of systems produced, we combine three different approaches to English-to-Czech MT in a sequential way as depicted in Figure 11.

Chimera consists of the following components:

TectoMT (Žabokrtský et al., 2008; Galuščáková et al., 2013) is a hybrid transfer-based MT system that processes English to obtain deep syntactic trees, transfers them to Czech trees using a statistical model and generates Czech sentence using a set of rules. Aside from the standard Czech morphological dictionary, TectoMT also includes a basic derivational dictionary so it can e.g. derive translations of adverbs even if only the corresponding adjectives were seen in parallel data. The target-side vocabulary of TectoMT is thus not restricted to observed items. Further gains can be expected from linguistically adequate handling of complex verbs.

The output of TectoMT *for the given input sentences*, i.e. for the source side of the test set, is included in the training data of the following step.

Moses (Koehn et al., 2007) is a standard phrase-based and hierarchical MT toolkit. It serves as the central component of Chimera, producing its single-best hypothesis. Trained on the large corpus CzEng (Bojar et al., 2012), data prepared by the organizers of WMT13[3] and about 3.6 gigaword of Czech news in the language model, our configuration is a strong system on its own.

The additional synthetic training data as provided by TectoMT allow Moses to produce words never seen in the parallel training data. The weights selecting whether to prefer translation of phrases from the parallel data or the output of

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

System	Automatic		Manual WMT Ranking		
	BLEU	TER	Appraise	MTurk	Total
CU-TECTOMT	14.7	0.741	0.455	0.491	0.476
CU-BOJAR	**20.1**	0.696	0.637	**0.555**	**0.580**
CU-DEPFIX = Chimera	20.0	**0.693**	**0.664**	0.542	0.578
Plain Moses	19.5	0.713	–	–	–
Google Translate	19.4	0.720	0.618	0.526	0.562

Table 6. Results of Chimera, an English-to-Czech system that combines TectoMT, Moses and Depfix in a sequential way. Best scores in bold.

TectoMT are automatically optimized on a heldout dataset. More details on the tuning are available in Bojar et al. (2013c).

Depfix (Rosa et al., 2012) is a system for correcting errors in (esp. phrase-based) MT outputs. As all components of Chimera, it makes use of the original English sentence to reduce the problem of error cummulation. Depfix parses both Moses output and the original input, fixing the former on the basis of the latter if needed. Hand-written rules then specify which words in the dependency tree need to agree in grammatical categories such as case or gender with the target language, or match with the source in number or negation.

Table 6 presents the results of Chimera on WMT13 test set using two automatic measures (BLEU and TER) and the official manual ranking. The manual ranking was performed by two groups of people: researchers and their colleagues (labelled "Appraise" after the annotation frontend) and random paid annotators using Amazon Mechanical Turk crowdsourcing platform ("MTurk" in the table). Note that the actual annotation interface was identical for both of the groups. The official results of WMT13 (Bojar et al., 2013a) do not distinguish between the annotator groups and we list the scores in the column "Total".

We submitted all the three steps of Chimera as independent systems to the evaluation. A big jump in quality comes with the powerful statistical Moses. Measured by both automatic metrics, TectoMT is an important component of the mix, raising BLEU of 19.5 ("Plain Moses", i.e. the same setup except without access to the output of TectoMT) to 20.0 (CU-BOJAR).

The impact of Depfix seems less clear. Since Depfix alters only a few words in each sentence, it is not surprising that the automatic scores do not change much. Depfix also does not use any language models, losing a little in the n-gram-based evaluation by BLEU.

The most interesting is the discrepancy in manual scores of Turkers and researchers. Turkers prefer the outputs without Depfix while researchers clearly appreciate corrected outputs better (0.637 vs. 0.664). We speculate that Turkers, rewarded for each submitted item of annotation, hurried up and evaluated the outputs more superficially. The critical little differences corrected by Depfix (most notably lost negation) may have often remained unnoticed. As Bojar et al. (2013a) reports in Table 3, the inter-annotator agreement among Turkers evaluating Czech was exceptionally low this year, reaching only κ of 0.075 compared to 0.408 for researchers or to 0.25 which is the average κ of Turkers across all language pairs.

Given the low reliability of Turker judgements, we conclude that Depfix does play an important role in the final translation quality.

Overall, Chimera outperfomed Google Translate in both automatic and manual scores, ranking first among the English-to-Czech systems.

3.8. Fully Automated Research in MT Not Feasible

As apparent from the previous sections, MT systems are very complex cascades of processing steps, each of which is influenced by various parameters. Finding the right configuration of these parameters is critical for system performance. Some of the parameters have the form of a real value or a vector of real values. For these, we follow the common practice and use a variant of a grid search. However, many settings are categorical or stand outside of the standard automatic search: for instance, there are several different algorithms available for the search itself.

We see research as the search for the best (design and) configuration of a complex system. We developed tools that support this vision and allow for manual or even fully automatic search in the space of MT system configuration.

The core of our tools is EMAN (Bojar and Tamchyna, 2013), a generic experiment manager that promotes to represent experiments as acyclic graphs of basic processing steps. EMAN facilitates the creation of the individual steps and most importantly the derivation of steps and whole experiments by altering existing ones. For MT, our EMAN steps correspond to tasks such as word alignment, extraction of translation or language models, model optimization or translation of unseen texts. EMAN includes an assistant for organizing the obtained results and makes it easy to handle dozens or hundreds of experiments without losing focus.

EMAN also served as the basis of *fully automatic* search for the best MT system. With EMAN, it is easy to examine the full Cartesian product of various settings or to navigate in this space using grid search or e.g. genetic algorithms. In Tamchyna and Bojar (2013), we however document that the domain of machine translation is too complex to be examined automatically, even when restricted to the single model of phrase-based translation and one particular language pair. There are two reasons of this infeasibility:

Too large space of configurations. In our experiments, we included source and target-side features from the morphological, surface and deep syntactic analysis of the sentence. Picking which features to choose and how to use them allows for more than a thousand possible configurations even in a very restricted experiment design. Each of these configurations would require new extraction of translation tables and a new model optimization, which is quite computationally expensive.

Too imprecise evaluation. The main problem, however, is the impossibility to tell better and worse systems apart, especially if the difference between them is not very big. The optimization of a given configuration is a randomized process leading to various results. Multiple runs of the examined configurations led to scores so different that it was possible to completely reverse the ranking of some of the configurations. In short, the resulting scores are similar and their variance is high.

To conclude, it is not possible to find the best system configuration fully automatically, but EMAN can at least support researchers in their semi-automatic examination of the space.

4. Conclusion

In the present article, we summarized the results of the project CzechMATE, which focused on statistical machine translation into Czech. Four different source languages were examined: English, German, Spanish and French.

Throughout the project, we contributed to the state of the art in several ways, ranging from techniques of manual and automatic MT evaluation over comparison of direct translation and pivoting through English (e.g. translation from German to Czech via English, where pairwise parallel data are easier to obtain), translation of words broken into unsupervised morphemes to experiments with handling named entities. We conclude by describing our combined MT system called Chimera that obtained very high ranking in the WMT 2013 shared task and by attempts to (fully) automate research in MT.

Acknowledgements

The work on this project was supported by the grant P406/11/1499 of the Czech Science Foundation (GAČR).

Bibliography

Berka, Jan, Martin Černý, and Ondřej Bojar. Quiz-Based Evaluation of Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95:77–86, March 2011. ISSN 0032-6585.

Berka, Jan, Ondřej Bojar, Mark Fishel, Maja Popović, and Daniel Zeman. Tools for machine translation quality inspection. Technical Report TR-2013-50, Univerzita Karlova v Praze, MFF, ÚFAL, Praha, Czechia, 2013. URL `http://ufal.mff.cuni.cz/techrep/tr50.pdf`.

Birch, Alexandra, Barry Haddow, Ulrich Germann, Maria Nadejde, Christian Buck, and Philipp Koehn. The Feasibility of HMEANT as a Human MT Evaluation Metric. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 52–61, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2203.pdf.

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

Bojar, Ondřej and Aleš Tamchyna. The Design of Eman, an Experiment Manager. *The Prague Bulletin of Mathematical Linguistics*, 99:39–58, 2013. ISSN 0032-6585.

Bojar, Ondřej and Dekai Wu. Towards a Predicate-Argument Evaluation for MT. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 30–38, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-4204.

Bojar, Ondřej, Kamil Kos, and David Mareček. Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 86–91, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-2016.

Bojar, Ondřej, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3921–3928, Istanbul, Turkey, May 2012. ELRA, European Language Resources Association. ISBN 978-2-9517408-7-7.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August 2013a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2201.

Bojar, Ondřej, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. Scratching the Surface of Possible Translations. *LNCS*, 2013b. ISSN 0302-9743.

Bojar, Ondřej, Rudolf Rosa, and Aleš Tamchyna. Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 92–98, Sofia, Bulgaria, August 2013c. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2208.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W07/W07-0218.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, March 2009. Association for Computational Linguistics.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-1703. Revised August 2010.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-2103.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-3102.

Collins, Michael, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1219840.1219906.

Creutz, Mathias and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TLSP)*, 4(1)(3), 2007. URL http://dl.acm.org/citation.cfm?id=1187418.

Dreyer, Markus and Daniel Marcu. HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada, June 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N12-1017.

Fishel, Mark, Ondřej Bojar, Daniel Zeman, and Jan Berka. Automatic Translation Error Analysis. In *Text, Speech and Dialogue: 14th International Conference, TSD 2011*, volume LNAI 3658. Springer Verlag, September 2011. ISBN 3-540-28789-2.

Galuščáková, Petra, Martin Popel, and Ondřej Bojar. PhraseFix: Statistical post-editing of TectoMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 141–147, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2216.

Hálek, Ondřej, Rudolf Rosa, Aleš Tamchyna, and Ondřej Bojar. Named Entities from Wikipedia for Machine Translation. In Lopatková, Markéta, editor, *ITAT 2011 Information Technologies – Applications and Theory*, volume 788, pages 23–30, September 2011. ISBN 978-80-89557-02-8.

Hunt, James W. and M. Douglas McIlroy. An Algorithm for Differential File Comparison. Computing Science Technical Report 41, Bell Laboratories, June 1976.

Koehn, Philipp. Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In Frederking, Robert E. and Kathryn Taylor, editors, *AMTA*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124. Springer, 2004. ISBN 3-540-23300-8.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association* *for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073445.1073462.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Praha, Czechia, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P07-2045`.

Levenshtein, Vladimir Iosifovich. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, February 1966.

Lo, Chi-kiu and Dekai Wu. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 220–229, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1023`.

Lopatková, Markéta, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves*. Univerzita Karlova v Praze, Nakladatelství Karolinum, Praha, 2008. In cooperation with Karolína Skwarska, Eduard Bejček, Klára Hrstková, Michaela Nová and Miroslav Tichý.

Matusov, Evgeny, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose B. Marino, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. System Combination for Machine Translation of Spoken and Written Language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237, September 2008.

Och, Franz Josef. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075096.1075117.

Och, Franz Josef and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Panevová, Jarmila. Valency Frames and the Meaning of the Sentence. In Luelsdorff, Ph. L., editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243, Amsterdam-Philadelphia, 1994. John Benjamins.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135.

Parker, Robert, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword fifth edition, LDC2011T07, June 2011. URL `http://catalog.ldc.upenn.edu/LDC2011T07`.

Popel, Martin and Zdeněk Žabokrtský. TectoMT: Modular NLP Framework. In Loftsson, Hrafn, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg, 2010. Iceland Centre for Language Technology (ICLT), Springer. ISBN 978-3-642-14769-2.

Rosa, Rudolf, David Mareček, and Ondřej Dušek. DEPFIX: A System for Automatic Correction of Czech MT Outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 362–368, Montréal, Canada, June 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W12-3146`.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands, 1986.

Stolcke, Andreas. SRILM – an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002.

Tamchyna, Aleš and Ondřej Bojar. No Free Lunch in Factored Phrase-Based Machine Translation. In *Proc. of CICLing 2013*, volume 7817 of *LNCS*, pages 210–223, Samos, Greece, 2013. Springer-Verlag.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genova, Italy, May 2006.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, OH, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-09-1.

Zeman, Daniel. Data issues of the multilingual translation matrix. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 395–400, Montréal, Canada, 2012. Association for Computational Linguistics. ISBN 978-1-937284-20-6.

Zeman, Daniel, Mark Fishel, Jan Berka, and Ondřej Bojar. Addicter: What is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96:79–88, 2011.

Address for correspondence:
Daniel Zeman
`zeman@ufal.mff.cuni.cz`
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
CZ-11800 Praha, Czechia