Approximating a Deep-Syntactic Metric for MT Evaluation and Tuning*

Matouš Macháček and Ondřej Bojar
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague
{bojar,machacek}@ufal.mff.cuni.cz

Abstract

SemPOS is an automatic metric of machine translation quality for Czech and English focused on content words. It correlates well with human judgments but it is computationally costly and hard to adapt to other languages because it relies on a deep-syntactic analysis of the system output and the reference. To remedy this, we attempt at approximating SemPOS using only tagger output and a few heuristics. At a little expense in correlation to human judgments, we can evaluate MT systems much faster. Additionally, we describe our submission to the Tunable Metrics Task in WMT11.

1 Introduction

SemPOS metric for machine translation quality was introduced by Kos and Bojar (2009). It is inspired by a set of metrics relying on various linguistic features on syntactic and semantic level introduced by Giménez and Márquez (2007). One of their best performing metrics was Semantic role overlapping: the candidate and the reference translation are represented as bags of words and their semantic roles. The similarity between the candidate and the reference is calculated using a general similarity measure called Overlapping. The formal definition may be found in Section 4.

Instead of semantic role labels (not available for Czech), Kos and Bojar (2009) use TectoMT framework (Žabokrtský et al., 2008) to assign a semantic part of speech defined by Sgall et al. (1986). In addition they use t-lemmas (deep-syntactic lemmas) instead of surface word forms, which most importantly means that the metric considers *content words only*. In the following, we will use "sempos" to denote the semantic part of speech and "SemPOS" to denote the whole metric by Kos and Bojar (2009).

SemPOS correlates well with human judgments on system level, see Section 2 for a brief summary of how the correlation is computed. The main drawback of SemPOS is its computational cost because it requires full parsing up to the deep syntactic level to obtain t-lemmas and semposes. In Section 3 we propose four methods which approximate t-lemmas and semposes without the deep syntactic analysis. These methods require only part-of-speech tagging and therefore they are not only faster but also easier to adapt for other languages, not requiring more advanced linguistic tools.

Giménez and Márquez (2007) and Bojar et al. (2010) used different formulas to calculate the final overlapping.[1] In Section 4, we examine variations of the formula, adding one version of our own.

By combining one of the approximation techniques with one of the overlapping formulas, we ob-

* This work has been supported by the grants EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic), P406/10/P259, P406/11/1499, and MSM 0021620838.

[1] In fact, Giménez and Márquez (2007) released two versions of the paper. Both of them are nearly identical except for the formula for overlapping, so we asked the authors which of the two versions is correct. It turns out that Bojar et al. (2010), unaware of the second version of the paper, used the wrong one but still obtained good results. We therefore (re-)examine both versions.

Workshop	Filename	Sentences	To English from	To Czech from
WMT08	test2008	2000	de, es, fr	–
WMT08	nc-test2008	2028	cs	en
WMT08	newstest2008	2051	cs, de, es, fr	en
WMT09	newstest2009	2525	cs, de, es, fr	en
WMT10	newssyscombtest2010	2034	cs, de, es, fr	en

Table 1: Datasets used to evaluate the correlation with human judgments. For example: the testset "test2008" was used for translation to English from German, Spanish and French and it was not used for any translation to Czech.

tain a variant of our metric. The performance of the individual variants is reported in Section 5.

Section 6 is devoted to our submission to the Tunable Metrics shared task of the Sixth Workshop on Statistical Machine Translation (WMT11).

2 Method of Evaluation

Our primary objective is to create a good metric for automatic MT evaluation and possibly also tuning. We are not interested much in how close is our proposed approximation to the (automatic or manual) semposes and t-lemmas. Therefore, we evaluate only how well do our metrics (the pair of a chosen approximation and a chosen formula for the overlapping) correlate with human judgments.

2.1 Test Data

We use the data collected during three Workshops on Statistical Machine Translation: WMT08 (Callison-Burch et al., 2008), WMT09 (Callison-Burch et al., 2009) and WMT10 (Callison-Burch et al., 2010). So far, we study only Czech and English as the target languages. Our test sets are summarized in Table 1: we have four sets with Czech as the target language and 16 sets with English as the target language.

Each testset in each translation direction gives us for each sentence one hypothesis for each participating MT system. Human judges (repeatedly) ranked subsets of these hypotheses comparing at most 5 hypotheses at once and indicating some ordering of the hypotheses. The ordering may include ties. In WMT, these 5-fold rankings are interpreted as "simulated pairwise comparisons": all pairwise comparisons are extracted from each ranking. The HUMAN SCORE for each system is then the percentage of pairs where the system was ranked better or equal to its competitor.

2.2 Correlation with Human Judgments

For each metric we examine, the correlation to human judgments is calculated as follows: given one of the test sets (the hypotheses and reference translations), the examined metric provides a single-figure score for each system. We use Spearman's rank correlation coefficient between the human scores and the scores of the given metric to see how well the metric matches human judgments. Because tied ranks do not exist, the correlation coefficient is given by:

$$\rho = 1 - \frac{6 \sum\limits_{i} (p_i - q_i)^2}{n(n^2 - 1)} \quad (1)$$

Human scores across different test sets are not comparable, so we compute correlations for each test set separately and average them.

3 Approximations of SemPOS

We would like to obtain t-lemmas and semantic parts of speech without deep syntactic analysis, assuming only automatic tagging and lemmatization.

Except for one option (Section 3.4), we approximate t-lemmas simply by surface lemmas. For the majority of content words, this works perfectly, but there are several regular classes of words where the t-lemma differs. In such cases, the t-lemma usually consists of the lemma of the main content word and an auxiliary word that significantly changes the meaning of the content word. These are e.g. English phrasal verbs ("blow up" should have the t-lemma "blow_up") and Czech reflexive verbs ("smát_se").

The approximation of semantic part of speech deserves at least some minimal treatment. The following sections describe four variations of the approximation.

Morph. Tag	Sempos	Rel. Freq.
NN	n.denot	0.989
VBZ	v	0.766
VBN	v	0.953
JJ	adj.denot	0.975
NNP	n.denot	0.999
PRP	n.pron.def.pers	0.999
VB	v	0.875
VBP	v	0.663
VBD	v	0.810
WP	n.pron.indef	1.000
NNS	n.denot	0.996
JJR	adj.denot	0.813

Table 2: A sample of the mapping from English morphological tags to semposes, including the relative frequency, e.g. $\frac{\text{count(NN,n.denot)}}{\text{count(NN)}}$.

3.1 Sempos from Tag

We noticed that the morphological tag determines almost uniquely the semantic part of speech. We use the Czech-English sentence-parallel corpus CzEng (Bojar and Žabokrtský, 2009) to create a simple dictionary which maps morphological tags to most frequent semantic parts of speech. Some morphological tags belong almost always to auxiliary words which do not have a corresponding deep-syntactic node at all, so the t-lemma and sempos are not defined for them. We include these morphological tags in the dictionary and map them to a special sempos value "-". Ultimately, words with such sempos are not included in the overlapping at all.

Table 2 shows a sample of this dictionary. The high relative frequencies indicate that we are not losing too much of the accuracy: overall 93.6 % for English and 88.4 % for Czech on CzEng e-test.

The first approximation relies just on this (language-specific) dictionary. The input text is automatically tagged, the morphological tags are deterministically mapped to semposes using the dictionary and words where the mapping led to the special value of "-" are removed.

In the following, we label this method as APPROX.

3.2 Exclude Stop-Words

By definition, the deep syntactic layer we use represents more or less only content words. Most auxiliary words become only attributes of the deep-syntactic nodes and play no role in the overlapping between the hypothesis and the reference.

Our first approximation technique (Section 3.1) identifies auxiliary words only on the basis of the morphological tag. We attempt to refine the recall by excluding a certain number of most frequent words in each language. The frequency list was obtained from the Czech and English sides of the corpus CzEng. We choose the exact cut-off for stop-words in each language separately: 100 words in English and 220 words in Czech. See Section 5.1 below.

In the following, the method is called APPROX-STOPWORDS.

3.3 Restricting the Set of Examined Semposes

We noticed that the contribution of each sempos to the overall performance of the metric in terms of correlation to human judgments can differ a lot. One of the underlying reasons may be e.g. greater or lower tagging accuracy of certain word classes, another reason may be that translation errors in certain word classes may be more relevant for human judges of MT quality.

Tables 3 and 4 report the correlation to human judgments if only words in a given sempos are considered in the overlapping. Based on these observations, we assume that some sempos types raise the correlation of the overlapping with human judgments and some lower it. We therefore try one more variant of the approximation which considers only (language-specific) subset of semposes.

The approximation called APPROX-RESTR considers only these sempos tags in Czech: v, n.denot, adj.denot, n.pron.def.pers, n.pron.def.demon, adv.denot.ngrad.nneg, adv.denot.grad.nneg. The considered sempos tags for English are: v, n.denot, adj.denot, n.pron.indef.

3.4 T-lemma and Sempos Tagging

Our last approximation method differs a lot from the previous three approximations. We use the sequence labeling algorithm (Collins, 2002) as implemented in Featurama[2] to choose the t-lemma and sempos tag. The CzEng corpus (Bojar and Žabokrtský, 2009) serves to train two taggers: one for Czech and

[2] http://sourceforge.net/projects/featurama/

Tag	R. Fr.	Min.	Max.	Avg.
v	0.236	0.403	1.000	0.735
n.denot	0.506	0.189	1.000	0.728
adj.denot	0.124	0.264	0.964	0.720
n.pron.indef	0.019	0.224	1.000	0.639
n.quant.def	0.039	-0.084	0.893	0.495
n.pron.def.pers	0.068	-0.500	0.975	0.493
adv.pron.indef	0.005	-0.382	1.000	0.432
adv.denot.grad.neg	0.003	-1.000	0.904	0.413

Table 3: English semposes and their performance in terms of correlation with human judgments if only words of the given sempos in APPROX are checked for match with the reference. Averaged across all testsets. Overlapping CAP is used, see Section 4 below. Column R. Fr. reports relative frequency of each sempos in the testsets.

Tag	R. Fr.	Min.	Max.	Avg.
n.pron.def.pers	0.030	0.406	0.800	0.680
n.pron.def.demon	0.026	0.308	1.000	0.651
adj.denot	0.156	0.143	0.874	0.554
adv.denot.ngrad.nneg	0.047	0.291	0.800	0.451
adv.denot.grad.nneg	0.001	0.219	0.632	0.445
adj.quant.def	0.004	-0.029	0.800	0.393
n.denot.neg	0.037	0.029	0.736	0.391
adv.denot.grad.neg	0.018	-0.371	0.800	0.313
n.denot	0.432	-0.200	0.720	0.280
adv.pron.def	0.000	-0.185	0.894	0.262
adj.pron.def.demon	0.000	0.018	0.632	0.241
n.pron.indef	0.027	-0.200	0.423	0.112
adj.quant.grad	0.006	-0.225	0.316	0.079
v	0.180	-0.600	0.706	0.076
adj.quant.indef	0.002	-0.105	0.200	0.052
adv.denot.ngrad.neg	0.000	-0.883	0.775	0.000
n.quant.def	0.000	-0.800	0.713	-0.085

Table 4: Czech semposes. See Table 3 for explanation.

one for English. At each token, each of the taggers uses the word form, morphological tag and surface lemma (of the current and the previous two tokens) to choose one pair of t-lemma and sempos tag from a given set.

The set of possible t-lemma and sempos pairs is created as follows. At first the sempos set is obtained. We simply use all semposes being seen with the given morphological tag in the corpus. Then we find possible t-lemmas for each sempos. For most semposes we consider surface lemma as the only t-lemma. For the sempos tag "v" we also add t-lemmas composed of the surface lemma and some auxiliary word present in the sentence ("blow_up", "smát_se"). For some other sempos tags we add special t-lemmas for negation and personal pronouns ("#Neg", "#PersPron").

The overall accuracy of the tagger on the e-test is 97.9 % for English and 94.9 % for Czech, a better result on a harder task (t-lemmas also predicted) than the deterministic tagging in Section 3.1.

We call this approximation method TAGGER.

4 Variations of Overlapping

The original Overlapping defined by Giménez and Márquez (2007) is given in Equations 2 and 3:

$$\mathrm{O}(t) = \frac{\sum\limits_{w \in r_i} \mathrm{cnt}(w,t,c_i)}{\sum\limits_{w \in r_i \cup c_i} \max(\mathrm{cnt}(w,t,r_i), \mathrm{cnt}(w,t,c_i))} \tag{2}$$

where c_i and r_i denotes the candidate and reference translation of sentence i and $\mathrm{cnt}(w,t,s)$ denotes number of times t-lemma w of type (sempos) t appears in sentence s. For each sempos type t, Overlapping $\mathrm{O}(t)$ calculates the proportion of correctly translated items of type t. In this paper we will call this overlapping BOOST.

Equation 3 describes Overlapping of all types:

$$\mathrm{O}(*) = \frac{\sum\limits_{t \in T} \sum\limits_{w \in r_i} \mathrm{cnt}(w,t,c_i)}{\sum\limits_{t \in T} \sum\limits_{w \in r_i \cup c_i} \max(\mathrm{cnt}(w,t,r_i), \mathrm{cnt}(w,t,c_i))} \tag{3}$$

where T denotes the set of all sempos types. We will call this Overlapping BOOST-MICRO because it micro-averages the overlappings of individual sempos types.

Kos and Bojar (2009) used a slightly different Overlapping formula, denoted CAP in this paper:

$$\mathrm{O}(t) = \frac{\sum\limits_{w \in r_i} \min(\mathrm{cnt}(w,t,r_i), \mathrm{cnt}(w,t,c_i))}{\sum\limits_{w \in r_i} \mathrm{cnt}(w,t,r_i)} \tag{4}$$

To calculate Overlapping of all types, Kos and Bojar (2009) used ordinary macro-averaging. We call the method CAP-MACRO:

$$\mathrm{O}(*) = \frac{1}{|T|} \sum_{t \in T} \mathrm{O}(t) \tag{5}$$

The difference between micro- and macro-average is that in macro-average all types have

Reduction	Overlapping	Min.	Max.	Avg.
approx	cap-micro	0.409	1.000	0.804
orig	cap-macro	0.536	1.000	0.801
approx	cap-macro	0.420	1.000	0.799
approx-restr	cap-macro	0.476	1.000	0.798
tagger	cap-micro	0.409	1.000	0.790
orig	cap-micro	0.391	1.000	0.784
approx-restr	cap-micro	0.391	1.000	0.782
approx-stopwords	cap-micro	0.391	1.000	0.754
sempos-bleu		0.374	1.000	0.754
approx-stopwords	cap-macro	0.280	1.000	0.724
tagger	boost-micro	0.306	1.000	0.717
orig	boost-micro	0.324	1.000	0.711
approx-stopwords	boost-micro	0.133	1.000	0.697
approx-restr	boost-micro	0.126	1.000	0.688
approx	boost-micro	0.224	1.000	0.686
tagger	cap-macro	0.118	1.000	0.669
bleu		-0.143	1.000	0.628

Table 5: Metric correlations for English as a target language

Reduction	Overlapping	Min.	Max.	Avg.
approx-restr	cap-macro	0.400	0.800	0.608
tagger	cap-macro	0.143	0.800	0.428
orig	cap-macro	0.143	0.800	0.423
approx-restr	cap-micro	0.086	0.769	0.413
tagger	cap-micro	0.086	0.769	0.413
orig	cap-micro	0.086	0.741	0.406
approx-stopwords	cap-micro	0.086	0.790	0.368
approx	cap-micro	0.086	0.734	0.354
approx-stopwords	cap-macro	0.086	0.503	0.347
sempos-bleu		0.086	0.676	0.340
approx	cap-macro	0.086	0.469	0.338
tagger	boost-micro	0.086	0.664	0.337
bleu		0.029	0.490	0.279
orig	boost-micro	-0.200	0.692	0.273
approx-stopwords	boost-micro	-0.200	0.685	0.271
approx	boost-micro	-0.200	0.664	0.266
approx-restr	boost-micro	-0.200	0.664	0.266

Table 6: Metric correlations for Czech as a target language

the same weight regardless of count. For example O(n.denot) and O(adv.denot.grad.nneg) would have the same weight, however there are many more items of type n.denot than items of type adv.denot.grad.nneg (see Tables 3 and 4). We consider this unnatural and we suggest a new Overlapping formula CAP-MICRO:

$$\mathrm{O}(*) = \frac{\sum\limits_{t \in T} \sum\limits_{w \in r_i} \min(\mathrm{cnt}(w,t,r_i), \mathrm{cnt}(w,t,c_i))}{\sum\limits_{t \in T} \sum\limits_{w \in r_i} \mathrm{cnt}(w,t,r_i)} \tag{6}$$

In sum, we have three Overlappings which should be evaluated: BOOST-MICRO (Equation 3), CAP-MACRO (Equation 5), and CAP-MICRO (Equation 6).

5 Experiments

Table 5 shows the results for English as the target language. The first two columns denote the combination of an approximation method and an overlapping formula. For conciseness, we report only the minimum, maximum and average value among correlations of all test sets.

To compare metrics to original SemPOS, the table includes non-approximated variant ORIG where the t-lemmas and semposes are assigned by the TectoMT framework. For the purposes of comparison, we also report the correlations of BLEU (Papineni et al., 2002) and a linear combination of APPROX+CAP-MICRO and BLEU (even weights) under the name SEMPOS-BLEU since this metric was used in Tunable Metric Task (Section 6).

The best performing metric is the combination of approximation APPROX and overlapping CAP-MICRO. It actually slightly outperforms all non-approximated metrics. In general, the reductions APPROX and ORIG combined with CAP-MICRO or CAP-MACRO perform very well. Reductions APPROX-STOPWORDS and APPROX-RESTR do not improve on APPROX.

The TAGGER approximation correlates similarly to ORIG when micro-average is used.

Table 6 contains the results for Czech as the target language. The best performing metric for Czech is APPROX-RESTR together with CAP-MACRO. In general approximation APPROX-RESTR is better than APPROX-STOPWORDS which is slightly better than APPROX.

The success of overlapping CAP-MACRO in Czech is due to the higher contribution of less frequent semposes to the overall correlation. While in English the best correlating semposes are also very frequent (Table 3), this does not hold for Czech (Table 4). The underlying reasons have yet to be explained.

In both languages, the overlapping BOOST-MICRO has a very low correlation. We therefore consider this overlapping not suitable for any met-

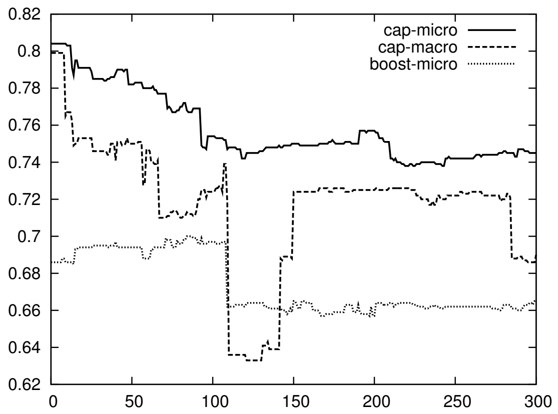

Figure 1: Correlation vs. the number of most frequent words which are thrown away for English. The big drop for lengths 109 and 110 is caused by the words 'who' and 'how'.

ric based on semposes.

On the other hand, most of the examined combinations are on average better than the baseline BLEU, sometimes by a very wide margin.

5.1 Dependency of Correlation on Stopwords List Length

We tried various stopwords list lengths for the approximation APPROX-STOPWORDS. Figure 5.1 shows the dependency of the correlation on stopwords list length for all overlappings in English. We see that the best correlation arises when no words are thrown away. One possible explanation is that auxiliary words are recognized by the morphological tag well enough anyway and stopwords lists remove also important content words, decreasing the overall accuracy of the overlapping.

6 Tunable Metric WMT11 Shared Task

The goal of the tunable metric task in WMT11 was to use the custom metric in MERT optimization (Och, 2003). The target language was English. We choose APPROX + CAP-MICRO since this combination correlates best with human judgments.

Based on the experience of Bojar and Kos (2010), we combine this metric with BLEU. In our opinion, the SemPOS metric and its variants alone are are good at comparing systems' outputs where sentence fluency has been already ensured. On the other hand, they fail in ranking sentences in n-best lists in MERT optimization because they observe only t-lemmas and don't penalize wrong morphological forms of words. We thus use BLEU to establish sentence fluency and our metrics to prefer sentences with correctly translated content words.

Weights		Devset scores	
BLEU	APPROX	BLEU	APPROX
1	0	0.246	0.546
0.75	0.25	0.242	0.584
0.5	0.5	0.229	0.594
0.25	0.75	0.215	0.602
0	1	0.025	0.631

Table 7: Results of MERT optimization. The last two columns contain metric scores of the last iteration of the MERT process with given combination weights.

We have tried several weights for the linear combination of BLEU and the chosen approximation. See Table 7 for details. We have submitted the variant with equal weights.

The preliminary results of manual evaluation (see the WMT11 overview paper) indicate that our system is fairly distinct from others: we won under the "> others" metric but we were the fifth of 8 systems in the official "≥ others" (the percentage of pairs where the system was ranked better or equal to its competitor).

7 Conclusions

We have introduced and evaluated several approximations of a deep-syntactic MT evaluation metric SEMPOS. This allows us to reduce the computational load by far, use only shallow tagging and still reach reasonable correlation scores.

For English, our combination of APPROX and CAP-MICRO performs even marginally better than the original SEMPOS. For Czech, it is APPROX-RESTR and TAGGER approximations with CAP-MACRO that outperform the original SEMPOS.

The applicability of these metrics (in link with BLEU) in model optimization was confirmed by the manual judgments for the Tunable Metrics Task. Our submission was surprisingly different from others: the best one in the score excluding ties and mediocre in the score where ties are rewarded.

References

Ondrej Bojar and Kamil Kos. 2010. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden, July. Association for Computational Linguistics.

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92. in print.

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*, pages 86–91, Uppsala, Sweden, July. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. 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. Association for Computational Linguistics. Revised August 2010.

Michael Collins. 2002. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Jesús Giménez and Lluís Márquez. 2007. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, June. Association for Computational Linguistics.

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

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

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

Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. 2008. 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. Association for Computational Linguistics.