Bulgarian-English and English-Bulgarian Machine Translation: System Design and Evaluation

Petya Osenova
Linguistic Modeling Department
IICT-BAS
petya@bultreebank.org

Kiril Simov
Linguistic Modeling Department
IICT-BAS
kivs@bultreebank.org

Abstract

The paper presents a deep factored machine translation (MT) system between English and Bulgarian languages in both directions. The MT system is hybrid. It consists of three main steps: (1) the source-language text is linguistically annotated, (2) it is translated to the target language with the Moses system, and (3) translation is post-processed with the help of the transferred linguistic annotation from the source text. Besides automatic evaluation we performed manual evaluation over a domain test suite of sentences demonstrating certain phenomena like imperatives, questions, etc.

1 Introduction

The paper describes a Deep Factored Moses system for Bulgarian-English MT in both directions, which includes transfer of linguistic knowledge from the source to the target language in the post-processing phase.

BG↔EN MT architecture presents a hybrid machine translation system that consists of three main steps. The source-language text is linguistically annotated, then translated with the Moses system to the target language and post-processed using the linguistic annotation projected from the source text. In the experiments we use four sets of parallel data — QTLeap corpus: Batch1 to Batch4 (Otegi et al., 2016) — and two versions of the translation systems: **BLMT** — baseline systems trained on pure parallel texts and **DFMT** — deep factored Bulgarian-English MT systems. In DFMT systems during the translation with the Moses system the word alignments are stored in order to be used for the projection of the linguistic analysis from the source text into the target one.

The structure of the paper is as follows: Section 2 presents the system architecture. Section 3 elaborates on the post-processing procedures. Section 4 presents the manual evaluation results. Section 5 outlines related work. Section 6 concludes the paper.

2 A Hybrid MT Architecture for DFMT

The hybrid architecture for DFMT for both language directions BG↔EN incorporates transfer of linguistic information from the source to the target language. It is depicted in Figure 1. The linguistic analysis for the source language (Analysis — column 1) is projected to a tokenized source text (Analysis — column 2). Then the Moses models (Moses) are applied to produce a target language translation. The translation alignment (Projection — column 1) is used for transferring the information to the corresponding tokens in the target language (Projection — column 2). The projected linguistic information interacts with the linguistic features of the tokens in the target text (for example the morphosyntactic features). The resulting annotation of the target text is used in the post-processing stage.

Note that the number of the tokens in the source and the target language might differ. The alignments can include many-to-many correspondences, not just one-to-one. Nevertheless, in practice about 80 % of the alignments are one-to-one or two-to-two tokens.

Here is an example of aligned texts annotated with morphosyntactic information of the

Proceedings of Recent Advances in Natural Language Processing, pages 561–568,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_073

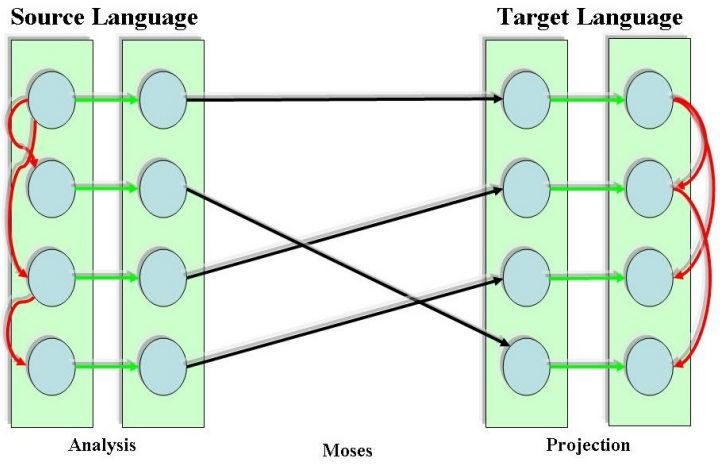

Figure 1: A hybrid architecture for transferring linguistic information from the source to the target text.

English[1] sentence "Place them in the midst of a pile of dirty, soccer kit." and its translation into Bulgarian[2]:

```
(place/VB them/PRP in/IN)
    =
   (postavyaneto/Ncnsd
    im/Ppetdp3;Ppetsp3;Pszt--3 v/R)
(the/DT midst/NN of/IN)
    = (razgar/Ncmsi na/R)
(a/DT pile/NN of/IN)
    = (kup/Ncmsi)
(dirty/JJ)    = (izmyrsyavam/Vpitf-r1s)
(,/,)         = (,/Punct)
(soccer/NN)   = (futbolni/A-pi)
(kit/NN)      = (komplekt/Ncmsi)
(./.)         = (./Punct)
```

From the alignment and rules for mappings between the two tagsets we could establish the following alignments on token level:

```
(them/PRP) = (im/Ppetdp3;Ppetsp3;Pszt--3)
(in/IN)    = (v/R)
(midst/NN) = (razgar/Ncmsi)
(of/IN)    = (na/R)
(pile/NN)  = (kup/Ncmsi)
(,/,)      = (,/Punct)
(soccer/NN)= (futbolni/A-pi)
(kit/NN)   = (komplekt/Ncmsi)
(./.)      = (./Punct)
```

Additionally, the alignment `(place/VB) = (postavyaneto/Ncnsd)` would be possible because the noun `(postavyaneto/Ncnsd)` is a deverbal noun, derived from a verb `(postavyam/Vpitf-r1s), to place`. To establish such an alignment we would need a derivational lexicon which however is not available to us. Thus, we do not consider this type of alignment. Likewise, the alignment between the English adjective `(dirty/JJ)` and the Bulgarian verb `(izmyrsyavam/Vpitf-r1s), to make dirty` would be also possible as much as "something to be dirty" could be a result from the action denoted by the verb. We consider such rules unreliable. Thus we do not use them. Using the alignments, we are able to transfer additional information like dependency links, word senses, etc. It is clear from the example that the transfer is only partial. The alignment `(soccer/NN) = (futbolni/A-pi)` is allowed when the English noun is a part of a compound. After the transfer of additional information, a set of rules for post processing is applied. For example, here a rule for agreement between the adjective `futbolni` and the noun `komplekt` has been applied.

For the EN→BG translation we extended the above architecture by adding an intermediate layer. Thus, the translation consists of two steps — see the modified architecture in Fig. 2. In the first step the English text is annotated with senses from the English WordNet. Then using the mapping between English WordNet and the Bulgarian Wordnet BTB-WN we substitute the English words with Bulgarian lemmas from the corresponding Bulgarian synsets. For each Bulgarian synset a representative lemma is preselected on the basis of a frequency list of Bulgarian lemmas that was

[1]For English, the tagset of Penn treebank is used: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

[2]For Bulgarian, the tagset of BulTreeBank was used: http://www.bultreebank.org/TechRep/BTB-TR03.pdf.

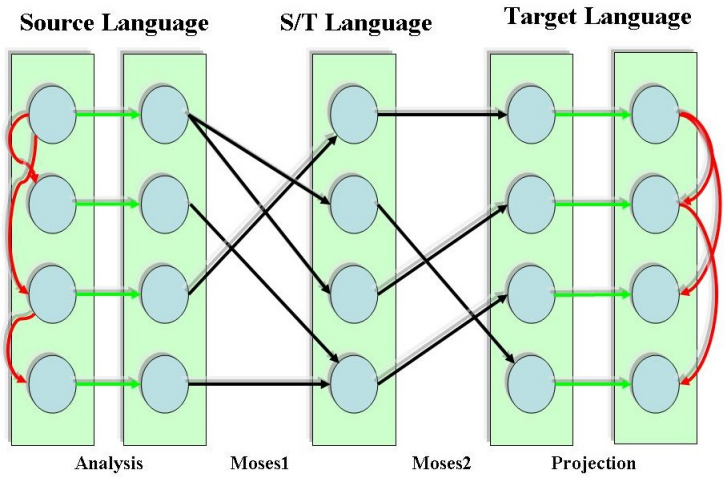

Figure 2: A hybrid architecture of EN→BG for transferring linguistic information from the source to the target language.

constructed over a corpus of 70 million words. The motivation for using the representative lemma in Bulgarian is our expectation for unification of the various synset ids with the similar translations in the target language. For example, the two concepts referred by *donor*: `wn30-10025730-n` ("person who makes a gift of property") and `wn30-10026058-n` ("a medical term denoting someone who gives blood or tissue or an organ to be used in another person") are very close to each other. They have the same translation in Bulgarian in both corresponding synsets: донор[3]. Here is an example of the performed processing:

English sentence:
This is real progress .

English sentence with factors:
this|this|dt is|be|vbz realen|real|jj napredyk|progress|nn .|.|.

Bulgarian sentence with factors (S/T Language Layer):
tova|tova|pd e|sym|vx realen|realen|a napredyk|napredyk|nc .|.|pu

Bulgarian sentence:
Tova e realen napredyk.

The intermediate layer is called **S/T Language** text. In addition we use the BLMT system and the alignment produced by the Moses System (Moses1) in order to substitute some of the functional words in source language with words in the target language. We selected this translation model for EN→BG because in our earlier experiments it performed slightly better than the phrase-based model. The architecture for EN→BG is depicted on Figure 2. Here the source language is analyzed linguistically, then the tokenized text is processed in two ways in order to produce the text for the Source/Target text (in the example above it corresponds to **English sentence with factors**). First, the replacements with the Bulgarian lemmas were performed on the basis of Word Sense annotation of the source text. Additionally, we translated the source text with the phrase-based Moses model (Moses1).

The linguistic analysis for the source language is projected to a tokenized source text; then Moses models (Moses1 and Moses2) are applied for producing a target language translation. The translation alignment is used for transferring the information to the corresponding tokens in the target language. Finally, the target linguistic annotation is used in the post-processing phase. It is important to keep in mind that the number of tokens in the S/T text is the same as in the source text. Thus, the analysis produced for the source text is easy to transfer to the S/T text. Then the actual translation was performed with factor-based Moses model (Moses2) where the alignment is used for the projection of the linguistic analysis over the source text.

[3]The corresponding Bulgarian synsets could contain other lemmas and thus for different domains and source language different representative lemmas will be more appropriate.

Our work on the projection of linguistic analysis from the source to the target text is similar to (Ramasamy et al., 2014) and (Mareček et al., 2011).

2.1 Analysis

For both directions the source-language linguistic annotation consists of tokenization, POS tagging, lemmatization, dependency parsing, shallow Minimal Recursion Semantics (MRS) annotation (elementary predicates and arguments) and word sense disambiguation. The analysis of English (tokenization, lemmatization, POS tagging and dependency parsing) was done with the CoreNLP tools.[4] The word sense disambiguation was done by the UKB tool.[5] For the analysis of Bulgarian, we trained Mate tools[6] on the Bulgarian treebank. The MRS structure rules were implemented in our own system.

2.2 Transfer

In the EN→BG system, we used a two-step translation strategy. **The first step**: (Moses1) was done using a phrase-based Moses model. We used the following parallel data: SETimes parallel corpus, LibreOffice parallel corpus, Bulgarian–English Dictionary aligned on wordform level, Microsoft product descriptions and Microsoft Terms. The texts were tokenized with available tokenizers for the corresponding languages. **The second step**: (Moses2) includes a factor-based Moses model which starts from a partially translated source language (S/T language). The training was performed on the same parallel data but processed as in the example given above on page 3. Both the source (**English sentence with factors**) and the target (**Bulgarian sentence with factors**) texts were processed with the corresponding language pipelines.

For the BG→EN system, we used the same parallel corpora and options as for the phrase-based model for EN→BG, but the whole transfer has been done in one step.

For the language model creation we relied on the SEtimes corpus, LibreOffice corpus, domain articles from Wikipedia, Microsoft data, and Europarl corpus. The tuning was done on Batch1 of the QTLeap corpus.

[4] http://stanfordnlp.github.io/CoreNLP/

[5] http://ixa2.si.ehu.es/ukb/

[6] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools.en.html

3 Post-processing

In the post processing step, a rule-based system was applied, based on linguistically-enhanced information. This information is projected from the source side with the help of the word alignments produced by Moses1 and Moses2. The projected information is the linguistic knowledge in the form of MRS-based elementary predicates, labeled dependencies, word senses (synset ids from WordNet) and POS tags in the source language.

system	metric	BG→EN	EN→BG
BLMT	BLEU	18.54	20.30
DFMT	BLEU	**24.93**	**23.91**

Table 1: BLEU scores of BLMT and DFMT on translations of Batch4 part of the QTLeap Corpus.

It should be noted that the alignment between the source language to the S/T language and from the S/T language to the target language is not one-to-one. It generally maps sets of tokens from the source language to sets of tokens in the target language. Thus, as it was presented above in the example, the transfer of the linguistic information from the analysis of the source language to the target one is not straightforward. Here we apply heuristic rules. Thus, the transferred linguistic information is only partial. For the rules definition we also exploited the language resources and tools for the target language — a morphological lexicon, a lemmatizer, and a morphological generator.

Once the linguistic annotation is projected via the alignment, the post-processing can be applied. It includes various types of rules: morphological, syntactic and semantic. An example of a syntactic rule is the transformation of the English noun compounds into the appropriate syntactic structures in Bulgarian. The direct transfer is rare, since the NN compounds are not so frequent in Bulgarian. The combination in which the first noun is a Named Entity is the most frequent one in the domain data. In the case of a phrase with an adjective and a noun in Bulgarian, a morpho-

(A)

Source:	Go to the Contacts menu ≥ Advanced ≥ Back up contacts to file	*2 inst.*
BLMT:	Преход към контактите меню ≥ разширени ≥ подкрепи контакти до файл	*2 inst.*
DFMT:	Отидете на contacts меню > > разширени назад с контакти за file	*0 inst.*
Reference:	Отидете в менюто Contacts ≥ Advanced ≥ Back up contacts to file	

logical rule for agreement is applied.

Table 1 shows the results comparing BLMT and DFMT systems. It can be observed that considerable improvements have been achieved in both directions.

4 Manual Evaluation

Here we employ the methodology developed by (Avramidis et al., 2016) to measure the performance of our systems on certain domain specific phenomena like imperatives, questions, terminology, etc. The graphic on Figure 3 as well as the associated Table 2 show the following overall situation: DFMT cannot outperform the baseline BLMT on the mentioned phenomena. It, however, achieves some similar figures on separators, then on imperatives, verbs and terminology. The results drop mostly in the cases of compoundings and quotation marks. As it can be seen, although we have considerable improvements with respect to the automatic metrics, reported above in Table 1, the performance on selected language phenomena might indicate worse results. Thus, the objective evaluation is actually not a trivial task at all.

	#	BLMT	DFMT
imperatives	97	74%	68%
compounds	100	44%	35%
“>” separators	60	100%	98%
quotation marks	200	90%	69%
verbs	221	78%	73%
terminology	153	67%	60%
sum	831		
average		76%	56%

Table 2: Translation accuracy on manually evaluated sentences in Bulgarian focusing on particular phenomena. Test sets consist of hand-picked source sentences that include the respective phenomenon.

We present here some examples of the manual evaluation questionnaire in order to show how the evaluation was performed.

Example A depicts the analysis of the menu item. The source contains two instances of the separator. The BLMT system treats all separators correctly. DFMT system places the separators next to each other, so there is no correct instance.

Example B illustrates the translation of imperative forms. There are two correct instances in the source sentence in B. BFMT translates the second verb form from example B (select) correctly, but the first one (press) is translated in the form of a present tense.

Example C illustrates the translation of verbs. The results in this area are satisfactory. DFMT obtains the lower average value. The verb ‘duplicate’ is translated wrongly by the baseline system. DFMT does not translate it and also there is an English verb form in the Bulgarian sentence.

Example D illustrates the translation of terms. There is only one term in the example. It is translated in BLMT (although not grammatically correctly), but the verb in the sentence is translated like a noun. DFMT does not translate the term at all, but at the same time it translates the verb correctly.

5 Related Work

Our work is closely connected to the transfer-based MT models. Ideally, given the availability of two deep grammars for some language pair, we would be able to translate through the transfer of the deep representations. The transfer in this setting is usually implemented in the form of rewriting rules. For instance, in the Norwegian LOGON project (Oepen et al., 2004), the transfer rules were hand-written (Bond et al., 2005; Oepen et al., 2007), which involved a large amount of manual work. Graham and van Genabith (2008) and Graham et al. (2009) explored the automatic rule induction approach

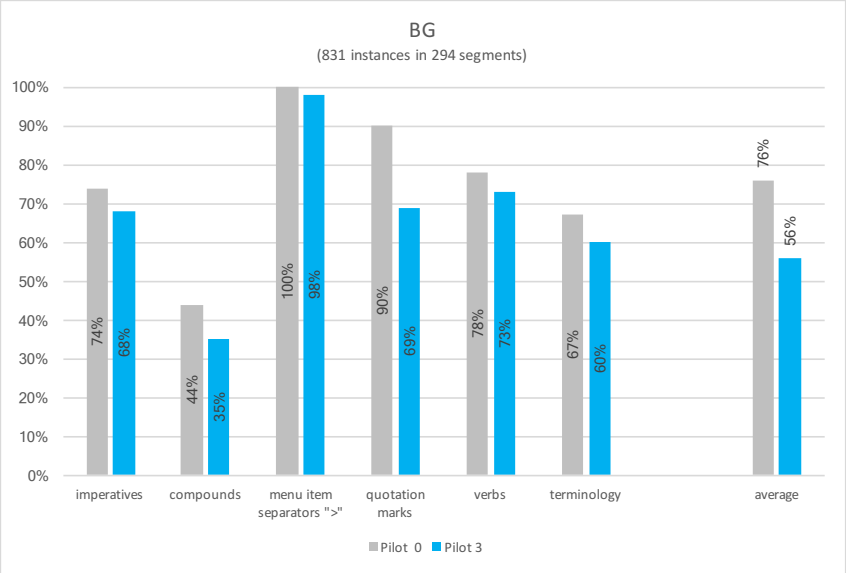

Figure 3: Manual evaluation results for Bulgarian

(B)

Source:	If you have WinRAR installed, press the right mouse button on the file and then select Extract here ...	*2 inst.*
BLMT:	Ако сте winrar инсталирани натискате десния бутон на мишката върху файла и отметнете добива тук ...	*1 inst.*
DFMT:	Ако имате winrar installed, натиснете десния бутон на мишката върху файла и след това изберете extract тук ...	*2 inst.*
Reference:	Ако имате инсталиран WinRAR, натиснете десния бутон на мишката върху файла и след това изберете Extract here ...	

in a transfer-based MT setting two Lexical Functional Grammars (LFGs), which was still restricted by the performance of both – the parser and the generator. Lack of robustness for target side generation is one of the main issues, when various ill-formed or fragmented structures come out after transfer. Oepen et al. (2007) use their generator to generate text fragments instead of full sentences, in order to increase the robustness.

However, since a real large-scale grammar for Bulgarian is still not available, we take an SMT system as our 'backbone' which robustly delivers some translation for any given input. Then, we incrementally augment SMT with deep linguistic knowledge. What we are doing is still along the lines of previous work utilizing deep grammars, but we build a transfer model over dependency parses.

Another stream of research is related to the TectoMT approach (Žabokrtský et al., 2008). The Prague Dependency Treebank (PDT)[7] is a Czech treebank, annotated in accordance to the linguistic theory of Functional Generative Description (P. Sgall and Panevova, 1986). The tectogrammatical layer[8] is the third layer of the PDT. It represents the syntactic-semantic interface, adding the functional dimension and collapsing the structural information, thus aiming at a more language-independent level of abstraction. The other two layers are the morphological and analytical ones. The morphological layer covers POS tags and lemmas. The analytical layer reflects the surface sentence structure. The tectogrammatical annotation builds on the analytical level. It presents the deep semantic structure of the sentence. The tectogrammatical level representation contains all the in-

[7] https://ufal.mff.cuni.cz/pdt2.0/

[8] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch01.html

(C)

Source:	Press and hold the Alt key and then click the color you want to duplicate.	*5 inst.*
BLMT:	Натиснете и задръжте върху Alt и щракнете върху цвета, който желаете да дублира.	*4 inst.*
DFMT:	Натиснете и задръжте клавиша Алт ключ и после щракнете върху цвят искате да duplicate.	*4 inst.*
Reference:	Натиснете и задръжте клавиша Alt и след това кликнете върху цвета, който искате да използвате.	

(D)

Source:	In the terminal, type ”netstat-a”.	*1 inst.*
BLMT:	В терминал, тип "netstat-a".	*1 inst.*
DFMT:	В terminal, въведете "netstat-a ".	*0 inst.*
Reference:	В терминала напишете „netstat-a".	

formation necessary for translating the tectogrammatical representation into the lower levels, as well as for its interpretation in the sense of intentional semantics.

In contrast to the analytical level, the tectogrammatical level highlights the functional dimension (such as the semantic roles *Actor, Patient, Addressee*, etc.). It abstracts away from the synsemantic (functional) parts-of-speech (prepositions, conjunctions, etc.) in the dependency trees, thus focusing on the autosemantic words (nouns, verbs, etc.). The structural information is not lost, but just “collapsed” into the content words representations. In this way, a more abstract level of language representation is achieved, which then is used for the transfer step within the MT systems. Our approach generally follows the ideas behind the tectogrammatical approach in the sense that we also abstract over the sentence structures, focusing on the content words. However, we use MRS as a logical form for sentence representations (elementary predicates, combination rules, etc.).

We also build on the previous language model translation experience described in (Wang et al., 2012a) and (Wang et al., 2012b), enhancing the factored architecture with a more elaborated transfer step and with a more linguistically-aware post-processing step. However, while in the above-mentioned publications only BG→EN translation was explored, in this paper also the EN→BG direction is presented.

6 Conclusions

The presented MT system contains several improvements over the baseline variants of the BG↔EN systems, developed by us. These include: improved knowledge graphs for WSD; extension of the parallel data with aligned terminology and multi-word expressions; rules for generation of shallow MRS structures, rules for transfer of linguistic information from source to target text and post-processing rules that were implemented manually.

The manual evaluation performed on selected language and domain phenomena shows, that even though the automatic evaluation might show improvements, the performance over the selected domain language phenomena might not be good or might even drop.

Our goal in the future is to develop a bigger test suite of phenomena for the language pair of Bulgarian and English. Another improvement we envisage is to add to the test suite a procedure that would allow automatic checking when the translations do not reflect the selected phenomena. Even if done only partially, it would save the manual work during this type of evaluation.

Acknowledgements

This research has received partial funding from the EC’s FP7 under grant agreement number 610516: “QTLeap: Quality Translation by Deep Language Engineering Approaches”. We are grateful to the anonymous reviewers for their remarks, comments, and suggestions.

References

Eleftherios Avramidis, Aljoscha Burchardt, Vivien Macketanz, and Ankit Srivastava. 2016. DFKI's system for wmt16 it-domain task, including analysis of systematic errors. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 415–422.

Francis Bond, Stephan Oepen, Melanie Siegel, Ann Copestake, and Dan Flickinger. 2005. Open source machine translation with DELPH-IN. In *Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit*. pages 15 – 22.

Yvette Graham, Anton Bryl, and Josef van Genabith. 2009. F-structure transfer-based statistical machine translation. In *Proceedings of the Lexical Functional Grammar Conference*. CSLI Publications, Stanford University, USA, Cambridge, UK., pages 317–328.

Yvette Graham and Josef van Genabith. 2008. Packed rules for automatic transfer-rule induction. In *Proceedings of the European Association of Machine Translation Conference (EAMT 2008)*. Hamburg, Germany, pages 57–65.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011. Two-step translation with grammatical post-processing. In Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan, editors, *Proceedings of the Sixth WMT*. University of Edinburgh, ACL, Edinburgh, UK, pages 426–432.

Stephan Oepen, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, , and Victoria Rosén. 2004. Som å kapp-ete med trollet? towards MRS-based norwegian to english machine translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Baltimore, MD.

Stephan Oepen, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. 2007. Towards hybrid quality-oriented machine translation — on linguistics and probabilities in MT. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*. Skovde, Sweden.

Arantxa Otegi, Nora Aranberri, António Branco, Jan Hajic, Martin Popel, Kiril Simov, Eneko Agirre, Petya Osenova, Rita Pereira, João Silva, and Steven Neale. 2016. Qtleap wsd/ned corpora: Semantic annotation of parallel corpora in six languages. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

E. Hajicova P. Sgall and J. Panevova. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects.*. Dordrecht: Reidel Publishing Company and Prague: Academia.

Loganathan Ramasamy, David Mareček, and Zdeněk Žabokrtský. 2014. Multilingual dependency parsing: Using machine translated texts instead of parallel corpora. *The Prague Bulletin of Mathematical Linguistics* 102:93–104.

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

Rui Wang, Petya Osenova, and Kiril Simov. 2012a. Linguistically-augmented bulgarian-to- english statistical machine translation model. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), EACL 2012*. pages 119—128.

Rui Wang, Petya Osenova, and Kiril Simov. 2012b. Linguistically-enriched models for bulgarian-to-english machine translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST-6, 2012*. pages 10—19.