A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 159–170

A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines

Aljoscha Burchardt,[a] Vivien Macketanz,[a] Jon Dehdari,[a] Georg Heigold,[a] Jan-Thorsten Peter,[b] Philip Williams[c]

[a] German Research Center for Artificial Intelligence (DFKI)
[b] RWTH Aachen University
[c] University of Edinburgh

Abstract

In this paper, we report an analysis of the strengths and weaknesses of several Machine Translation (MT) engines implementing the three most widely used paradigms. The analysis is based on a manually built test suite that comprises a large range of linguistic phenomena. Two main observations are on the one hand the striking improvement of an commercial online system when turning from a phrase-based to a neural engine and on the other hand that the successful translations of neural MT systems sometimes bear resemblance with the translations of a rule-based MT system.

1. Introduction

Test suites are a familiar tool in NLP in areas such as grammar checking, where one may wish to ensure that a parser is able to analyse certain sentences correctly or test the parser after changes to see if it still behaves in the expected way. In contrast to a "real-life" corpus the input in a test suite may well be made-up or edited to isolate and illustrate issues.

Apart from several singular attempts (King and Falkedal, 1990; Isahara, 1995; Koh et al., 2001, etc.) broadly-defined test suites have not generally been used in MT research. One of the reasons for this might be the fear that the performance of statistical MT systems depends so much on the particular input data, parameter settings, etc., that final conclusions about the errors they make, particularly about the different

 Corresponding author: aljoscha.burchardt@dfki.de
Cite as: Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, Philip Williams. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 159–170. doi: 10.1515/pralin-2017-0017.

reasons (e.g., length of n-grams, missing training examples), are difficult to obtain. A related concern is that statistical MT systems are designed to maximise scores on test corpora that are comparable to the training/tuning corpora and that it is therefore unreliable to test these systems in different settings. While these concerns may hold for systems trained on very narrowly-defined domains, genres, and topics (such as biomedical patent abstracts), in fact many systems are trained on large amounts of data covering mixed sources and are expected to generalize to some degree.

A last reason might be that "correct" MT output cannot be specified in the same way as the output of other language processing tasks like parsing or fact extraction where the expected results can be more or less clearly defined. Due to the variation of language, ambiguity, etc., checking and evaluating MT output can be almost as difficult as the translation itself. Still, people have tried to automatically classify errors comparing MT output to reference translations or post-edited MT output using tools like Hjerson (Popovic, 2011).

In narrow domains there seems to be interest in detecting differences between systems and within the development of one system, e.g., in terms of verb-particle constructions (Schottmüller and Nivre, 2014) or pronouns (Guillou and Hardmeier, 2016). Bentivogli et al. (2016) performed a comparison of neural- with phrase-based MT systems on IWSLT data using a coarse-grained error typology. Neural systems have been found to make fewer morphological, lexical and word-order errors.

Below, we present a pioneering effort to address translation barriers in a systematic fashion. We are convinced that testing of system performance on error classes leads to insights that can guide future research and improvements of systems. By using test suites, MT developers will be able to see how their systems perform compared to scenarios that are likely to lead to failure and can take corrective action.

This paper is structured as follows: After the general introduction (Section 1), Section 2 will briefly introduce the test suite we have used in the experiments reported in Section 3. Section 4 concludes the paper.

2. The Test Suite

The experiments reported below are based on a test suite for MT Quality we are currently building for the language pair English – German in the QT21 project. The test suite itself will be described in more detail in a future publication. In brief, it contains segments selected from various parallel corpora and drawn from other sources such as grammatical resources, e.g., the TSNLP Grammar Test Suite (Lehmann et al., 1996) and online lists of typical translation errors.

Each test sentence is annotated with the phenomenon category and the phenomenon it represents. An example showing these fields can be seen in Table 1 with the first column containing the source segment and the second and third column containing the phenomenon category and the phenomenon, respectively. The fourth column shows the translation given by the old Google Translate system and the last column

contains a post-edit of the MT output that is created by making as few changes as possible. In our latest version of the test suite, we have a collection of about 5,000 segments per language direction that are classified in about 15 categories (most of them similar in both language directions) and about 120 phenomena (many of them similar but also some differing, as they are language-specific). Depending on the nature of the phenomenon, each is represented by at least 20 test segments in order to guarantee for a balanced test set. The categories cover a wide range of different grammatical aspects that might or might not lead to translation difficulties for a MT system. Currently, we are still in the process of optimising our test segments and working on an automatic solution for the evaluation.

Source	**Phenomenon Category**	**Phenomenon**	**Target (raw)**	**Target (edited)**
Lena machte sich früh vom Acker.	MWE	Idiom	Lena [left the field early].	Lena left early.
Lisa hat Lasagne gemacht, sie ist schon im Ofen.	Non-verbal agreement	Coreference	Lisa has made lasagne, [she] is already in the oven.	Lisa has made lasagna, it is already in the oven.
Ich habe der Frau das Buch gegeben.	Verb tense/ aspect/ mood	Ditransitive - perfect	I [have] the woman of the Book.	I have given the woman the book.

Table 1. Example test suite entries German→English (simplified for display purposes).

For the experiments presented here, we have used a preliminary version of our test suite (ca. 800 items per language direction, to a large extent verb paradigms) to include the changes of Google Translate which has recently been switched from a phrase-based to neural approach according to the companies' publications. There are more than 100 different linguistic phenomena that we investigated in this version of the test suite in each language direction. In this preliminary version, the number of instances reported in the experiments below strongly varies among the categories (as well as between the languages).

3. Evaluating PBMT, NMT, and RBMT Engines and an Online System

3.1. System Description

We have evaluated several engines from leading machine translation research groups and a commercial rule-based system on the basis of the very same test suite version to be able to compare performance with the leading online system that has recently switched to a neural model. We included a number of different NMT systems with different properties and levels of sophistication to shed light on how these

new types of systems perform on the different kinds of phenomena. Below, we will briefly describe the systems.

O-PBMT Old version of Google Translate (web interface, Feb. 2016).

O-NMT New version of Google Translate (web interface, Nov. 2016).

OS-PBMT Open-source phrase-based system that primarily uses a default configuration to serve as a baseline. This includes a 5-gram modified Kneser-Ney language model, `mkcls` and MGiza for alignment, GDFA phrase extraction with a maximum phrase length of five, `msd-bidi-fe` lexical reordering, and the Moses decoder (Koehn et al., 2007). The WMT'16 data was Moses-tokenized and normalized, truecased, and deduplicated.

DFKI-NMT Barebone neural system from DFKI. The MT engine is based on the encoder-decoder neural architecture with attention. The model was trained on the respective parallel WMT'16 data.

ED-NMT Neural system from U Edinburgh. This MT engine is the top-ranked system that was submitted to the WMT '16 news translation task (Sennrich et al., 2016). The system was built using the Nematus toolkit.[1] Among other features, it uses byte-pair encoding (BPE) to split the vocabulary into subword units, uses additional parallel data generated by back-translation, uses an ensemble of four epochs (of the same training run), and uses a reversed right-to-left model to rescore n-best output.

RWTH-NMT NMT-system from RWTH (only used for German – English experiments). This system is equal to the ensemble out of 8 NMT systems optimized on TEDX used in the (Peter et al., 2016) campaign. The eight networks used make use of subwords units and are finetuned to perform well on the IWSLT 2016 MSLT German to English task.

RBMT Commercial rule-based system Lucy (Alonso and Thurmair, 2003).

3.2. Evaluation Procedure

In order to evaluate a system's performance on the categories in the test suite, we concentrate solely on the phenomenon in the respective sentence and disregard other errors. This means that we have to determine whether a translation error is linked to the phenomenon under examination or if it is independent from the phenomenon. If the former is the case, the segment will be validated as incorrect. If, however, the error in the translation can not be traced back to the phenomenon, the segment will be counted as correct.

Currently, the system outputs are being automatically compared to a "reference translation" which is, in fact, a post-edit of the O-PBMT output as those were the very first translations to be generated and evaluated when we started building the test suite (see description of the test suite in Section 2 and Table 1). In a second step,

[1] `https://github.com/rsennrich/nematus`

	#	O-PBMT	O-NMT	RBMT	OS-PBMT	DFKI-NMT	RWTH-NMT	ED-NMT
Ambiguity	17	12%	35%	**42%**	24%	35%	12%	35%
Composition	11	27%	**73%**	55%	27%	45%	45%	**73%**
Function words	19	5%	**68%**	21%	11%	26%	**68%**	42%
LDD & interrogative	66	12%	**79%**	62%	21%	36%	55%	52%
MWE	42	14%	**36%**	7%	21%	10%	12%	19%
NE & terminology	25	48%	48%	40%	**52%**	40%	48%	40%
Subordination	36	22%	**58%**	50%	31%	47%	42%	31%
Verb tense/aspect/mood	529	59%	80%	**91%**	52%	53%	74%	63%
Verb valency	32	16%	**50%**	44%	13%	47%	38%	**50%**
Sum	777	358	567	583	337	367	490	435
Average		46%	73%	**75%**	43%	47%	63%	56%

Table 2. Results of German – English translations. Boldface indicates best system(s) on each category (row).

all the translations that do not match the "reference" are manually evaluated by a professional linguist since the translations might be very different from the O-PBMT post-edit but nevertheless correct. As this is a very time-consuming process, we are currently working on automating this evaluation process by providing regular expressions for various possible translation outputs – naturally, only focusing on the phenomenon under investigation.

We refrain from creating an independent reference as we think that generating the regular expressions that focus solely on the phenomena instead is the more sophisticated solution in this context. As a consequence, we cannot compute automatic scores like BLEU. We do not see this as a disadvantage as with the test suite we want to focus rather on gaining insights about the nature of translations than on how well translations match a certain reference.

3.3. Results German – English

Table 2 shows the results for the translations from German to English from the different systems on the categories. The second column in the table ("#") contains the number of instances per category. As the distribution of examples per category in this old version of our test suite was very unbalanced with some categories having only very few examples, some more categories we tested were excluded from the analysis we present here.

Before we discuss the results, we want to point out that the selection of phenomena and the number of instances used here is not representative of their occurrence in

corpora. Consequently, it can not be our goal to find out which of the systems is the globally "best" or winning system. Our goal is to check and illustrate the strengths and weaknesses of system (types) with respect to the range of phenomena we cover with this version of the test suite. Using this evaluation approach, researchers and system developers ideally can form hypotheses about the reasons why certain errors happen (systematically) and can come up with a prioritised strategy for improving the systems. Our ultimate goal is to represent all phenomena relevant for translation in our test suite.

Coming to the analysis, it is first of all striking how much better the neural version of Google Translate (O-NMT) is as compared to its previous phrase-based version (O-PBMT). Interestingly, the O-NMT and the RBMT – two very different approaches – are the best-performing systems on average, achieving almost the same amount of correct translations on average, i.e., 73%, resp. 75%, but looking at the scores of the categories reveals that the performance of the two systems regarding the categories is in fact very diverse. While the O-NMT system is the most-frequent best-performing system per phenomenon, as it is best on composition, function words, long distance dependency (LDD) & interrogative, multi-word expressions (MWE), subordination and verb valency, the RBMT is only the best system on ambiguity[2] and verb tense/aspect/mood. The high number of instances of the latter category leads to the high average score of the RBMT system, as verb paradigms are part of the linguistic information RBMT systems are based on.

The OS-PMBT reaches the lowest average score, but it is nevertheless the best-performing system on named entities (NE) & terminology. The DFKI-NMT system reaches a higher average score than the PBMT system (four percentage points more). The RWTH-NMT is (along with the O-NMT) the best-performing system on function words. On average it reaches 63% of correct translations. The ED-NMT outrules (also along with the O-NMT) the other systems on composition and verb valency and reaches 56% correct translations on average.

In order to see if we find some interesting correlations that might serve as a preview for more extensive analyses with a more solid and balanced amount of test segments in the future, we have calculated Pearson's coefficient over the phenomenon counts (being aware that we are dealing with very small numbers here). As the correlations for the direction English – German were higher and for space reasons, we will show the numbers only for the other direction in the following Subsection to give an indication about possible future work.

One general impression that will also be supported by the examples below is that NMT seems to learn some capabilities that the RBMT system has. It may lead to the speculation that NMT indeed learns something like the rules of the language. This, however, needs more intensive investigation. Another interesting observation is that

[2]The good performance of RBMT on ambiguity can be explained by the very small number of items and it is more or less accidental that the preferred readings were the ones the RBMT has coded in its lexicon.

the RWTH-NMT system has a lower overall correlation with the other NMT systems. This might be because it has also been trained and optimised on transcripts of spoken language as opposed to the other systems trained solely on written language.

The following examples depict interesting findings from the analysis and comparison of the different systems. When a system created a correct output (on the respective category), the system's name is marked in boldface.

(1)	**Source**:	Warum hörte Herr Muschler mit dem Streichen auf?
	Reference:	Why did Mr. Muschler stop painting?
	O-PBMT:	Why heard Mr Muschler on with the strike?
	O-NMT:	Why did Mr. Muschler stop the strike?
	RBMT:	Why did Mr Muschler stop with the strike?
	OS-PBMT:	Why was Mr Muschler by scrapping on?
	DFKI-NMT:	Why did Mr Muschler listen to the rich?
	RWTH-NMT:	Why did Mr. Muschler listen to the stroke?
	ED-NMT:	Why did Mr. Muschler stop with the stump?

Example (1) contains a phrasal verb and belongs to the category composition. German phrasal verbs have the characteristics that their prefix might be separated from the verb and move to the end of the sentence in certain constructions, as it has happened in example (1) with the prefix *auf* being separated from the rest of the verb *hören*. The verb *aufhören* means *to stop*, but the verb *hören* without the prefix simply means *to listen*. Thus, phrasal verbs might pose translations barriers in MT when the system translates the verb separately not taking into account the prefix at the end of the sentence. The output of the O-PBMT, DFKI-NMT and RWTH-NMT indicates that this might have happened. The O-NMT, RBMT and the ED-NMT correctly translate the verb which could mean that more context (and thus, including the prefix *auf* at the end of the sentences) was taken into account for the generation of the output.

(2)	**Source**:	Warum macht der Tourist drei Fotos?
	Reference:	Why does the tourist take three fotos?
	O-PBMT:	Why does the tourist three fotos?
	O-NMT:	Why does the tourist make three fotos?
	RBMT:	Why does the tourist make three fotos?
	OS-PBMT:	Why does the tourist three fotos?
	DFKI-NMT:	Why does the tourist make three fotos?
	RWTH-NMT:	Why is the tourist taking three fotos?
	ED-NMT:	Why does the tourist make three fotos?

One of the phenomena in the category LDD & interrogative is wh-movement. It is for example involved in wh-questions, like in the sentence in (2). A wh-question in English is usually built with an auxiliary verb and a full verb, e.g., wh-word + *to*

have/to be/to do + full verb. In German on the other hand, an auxiliary verb is not necessarily needed. This fact might lead to translation difficulties, as can be seen in (2), where the O-PBMT and the OS-PBMT treat the verb *does* as a full verb instead of an auxiliary verb. All the other systems translate the question with two verbs, however, except for the RWTH-NMT, they all mistranslate *ein Foto machen* as *to make a foto* (literal translation) instead of *to take a foto*. Nevertheless, these translations count as correct, since they do contain an auxiliary verb + a full verb.

(3)
Source:	Die Arbeiter müssten in den sauren Apfel beißen.
Reference:	The workers would have to bite the bullet.
O-PBMT:	The workers would have to bite the bullet.
O-NMT:	The workers would have to bite into the acid apple.
RBMT:	The workers would have to bite in the acid apple.
OS-PBMT:	The workers would have to bite the bullet.
DFKI-NMT:	Workers would have to bite in the acid apple.
RWTH-NMT:	The workers would have to bite into the clean apple.
ED-NMT:	The workers would have to bite in the acidic apple.

Idioms are an interesting phenomenon within the category MWE. The meaning of an idiom in one language can not be transferred to another language by simply translating the separate words, as the meaning of these multi-word units goes beyond the meaning of the separate words. As a consequence, idioms have to be transferred to another language as a whole. For German <> English it is often the case that an idiom in one language can be transferred to another idiom in the other language. This is also the case in example (3). The German idiom *in den sauren Apfel beißen* can be translated as *to bite the bullet*. Only the two PBMT system correctly translate this idiom, the other systems all give a literal translation - with the RWTH-NMT translating *sauren* as *clean* instead of *acid(ic)* like the other systems, probably not knowing the word *sauren* and instead translating the similar word *sauberen*. This is one example where a phrase-based approach has a real advantage (if the phrase was in the training data).

(4)
Source:	Wie kann ich die Farbe, mit der ich arbeite, ändern?
Reference:	How can I change the color I am working with?
O-PBMT:	How can I change the color with which I work to change?
O-NMT:	How can I change the color with which I work?
RBMT:	How can I change the color with which I work?
OS-PBMT:	How can I change the colour, with whom i work, change?
DFKI-NMT:	How can I change the color I work with?
RWTH-NMT:	How can I change the color I work with?
ED-NMT:	How can I change the color I work with?

The sentence in (4) contains a relative clause which belongs to the category subordination. Relative clauses in English can, but do not have to contain a relative pronoun. The outputs in (4) show both properties. The O-PBMT and the OS-PBMT double the verb *change*, the remaining systems correctly translate the relative clause.

(5)

Source:	Ich hätte nicht lesen gedurft.
Reference:	I would not have been allowed to read.
O-PBMT:	I would not have been allowed to read.
O-NMT:	I should not have read.
RBMT:	I would not have been allowed to read.
OS-PBMT:	I would not have read gedurft.
DFKI-NMT:	I would not have been able to read.
RWTH-NMT:	I wouldn't have read.
ED-NMT:	I wouldn't have read.

Verb paradigms (verb tense/aspect/mood) make up about one third of the whole test suite. Example (5) shows a sentence with a negated modal verb, in the tense pluperfect subjunctive II. This is a quite complex construction, thus it is not surprising that only few systems correctly translate the sentence. As might be expected, one of them is the RBMT system. The second one is the O-PBMT. The neural version of this system on the other hand does not correctly produce the output.

3.4. Results English – German

The results for the English – German translations can be found in Table 3. For this language direction, only five systems were available instead of seven like for the other direction. As in the analysis for the other language direction, we excluded the categories that had too few instances from the table. Nevertheless, similarities between the categories of both language directions can be found.

As in the German – English translations, the RBMT system performs best of all systems on average, reaching 83%. It performs best of all systems on verb tense/aspect/mood and verb valency. The second-best system is – just like in the other language direction but with a greater distance (seven percentage points less on average, namely 76%) – the O-NMT. The O-NMT shows quite contrasting results on the different categories, compared to RBMT: it outrules (most of) the other systems on the remaining categories, i.e., on coordination & ellipsis, LDD & interrogative, MWE, NE & terminology, special verb types and subordination.

The third-best system on average is the ED-NMT system. It reaches an average of 61% correct translations. The other remaining NMT system, the barebone DFKI-NMT system, reaches 11 percentage points less on average than the ED-NMT, for it reaches 50%. But it outrules the other systems on subordination along with O-NMT. The system with the lowest average score is the previous version of Google Translate,

	#	O-PBMT	O-NMT	RBMT	DFKI-NMT	ED-NMT
Coordination & ellipsis	17	6%	**47%**	29%	24%	35%
LDD & interrogative	70	19%	**61%**	54%	41%	40%
MWE	42	21%	**29%**	19%	21%	26%
NE & terminology	20	25%	**80%**	40%	45%	65%
Special verb types	14	14%	**86%**	79%	29%	64%
Subordination	35	11%	**71%**	54%	**71%**	69%
Verb tense/aspect/mood	600	41%	82%	**96%**	53%	66%
Verb valency	22	36%	59%	**68%**	64%	59%
Sum	820	287	622	679	410	499
Average		35%	76%	**83%**	50%	61%

Table 3. Results of English – German translations. Boldface indicates best system(s) on each category (row).

Correlations	O-PBMT	O-NMT	RBMT	DFKI-NMT	ED-NMT
O-PBMT	1.00				
O-NMT	0.34	1.00			
RBMT	0.39	0.55	1.00		
DFKI-NMT	0.28	0.29	0.36	1.00	
ED-NMT	0.30	0.33	0.43	0.55	1.00

Table 4. Overall correlation of English – German systems

namely the O-PBMT. With 35% on average, it reaches less than half of the score of the O-NMT.

The results of the calculation of the Pearson's coefficient can be found in Table 4. Only categories with more than 25 observations had their correlation analysed. For the interpretation, we used a rule-of-thumb mentioned in the literature[3].

In the overall correlation, RBMT has a moderate correlation with O-NMT, which might be traced back to the fact that these are the two systems that correctly translate most of the test segments, compared to the other systems. The two neural systems, DFKI-NMT and ED-NMT, also have moderate correlations. All the other systems have weak correlation with each other.

Again, for the small and unbalanced numbers of samples, we do not want to put too much emphasis on the observations regarding correlations. This type of analysis might, however, become more informative in future work.

[3] http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r

4. Conclusions and Outlook

While the selection of test items/categories and even more the selection of examples we discussed provides a selective view on the performance of the system, we are convinced that this type of quantitative and qualitative evaluation provides valuable insights and ideas for improvement of the systems, e.g., by adding linguistic knowledge in one way or another. Two main observations we want to repeat here is the striking improvement of the commercial online system when turning from a phrase-based to a neural engine. A second observation is that the successful translations of some NMT systems often bear resemblance with the translations of the RBMT system. Hybrid combinations or pipelines where RBMT systems generate training material for NMT systems seem a promising future research direction to us.

While the extracted examples above give very interesting insights on the systems' performances on the categories, these are only more or less random spot tests. However, taking a close look at the separate phenomena at a larger scale and in more detail will lead to more general, systematic observations. This is what we aim to do with our current version of the test suite which is therefore much more extensive and systematic and therefore also allows for more general observations and more quantitative statements in future experiments.

Our ultimate goal is to automate the test suite testing. To this end, we are currently working on a method that is using regular expressions for automatically checking the output of engines on the test suite. The idea is to manually provide positive and negative tokens for each test item that can range from expected words in case of disambiguation up to, verbs and their prefixes with wild cards in between up to complete sentences in the case of verb paradigms.

Acknowledgements

This research is supported by the EC's Horizon 2020 research and innovation programme under grant agreements no. 645452 (QT21).

Bibliography

Alonso, Juan A and Gregor Thurmair. The Comprendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*. International Association for Machine Translation (IAMT), 2003.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus Phrase-Based Machine Translation Quality: a Case Study. *CoRR*, abs/1608.04631, 2016.

Guillou, Liane and Christian Hardmeier. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In Chair), Nicoletta Calzolari (Conference, 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 Interna-*

tional Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Isahara, Hitoshi. JEIDA's test-sets for quality evaluation of MT systems: Technical evaluation from the developer's point of view. In *Proceedings of the MT Summit V. Luxembourg*, 1995.

King, Margaret and Kirsten Falkedal. Using Test Suites in Evaluation of Machine Translation Systems. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2*, COLING '90, pages 211–216, Stroudsburg, PA, USA, 1990. Association for Computational Linguistics.

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, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Koh, Sungryong, Jinee Maeng, Ji-Young Lee, Young-Sook Chae, and Key-Sun Choi. A test suite for evaluation of English-to-Korean machine translation systems. In *Proceedings of the MT Summit VIII. Santiago de Compostela, Spain*, 2001.

Lehmann, Sabine, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. TSNLP - Test Suites for Natural Language Processing. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 711–716, 1996.

Peter, Jan-Thorsten, Andreas Guta, Nick Rossenbach, Miguel Graça, and Hermann Ney. The RWTH Aachen Machine Translation System for IWSLT 2016. In *International Workshop on Spoken Language Translation*, Seattle, USA, Dec. 2016.

Popovic, Maja. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–68, 10 2011.

Schottmüller, Nina and Joakim Nivre. Issues in Translating Verb-Particle Constructions from German to English. In *Proc. of the 10th Workshop on Multiword Expressions (MWE)*, pages 124–131, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT 16. *CoRR*, abs/1606.02891, 2016.

Address for correspondence:
Aljoscha Burchardt
aljoscha.burchardt@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany