A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018

Liane Guillou[1*] **Christian Hardmeier**[2*]
Ekaterina Lapshinova-Koltunski[3*] **Sharid Loáiciga**[4*]
[1]School of Informatics, University of Edinburgh
[2]Department of Linguistics and Philology, Uppsala University
[3]Department of Language Science and Technology, Saarland University
[4]CLASP, University of Gothenburg
lguillou@inf.ed.ac.uk christian.hardmeier@lingfil.uu.se
e.lapshinova@mx.uni-saarland.de sharid.loaiciga@gu.se

Abstract

We evaluate the output of 16 English-to-German MT systems with respect to the translation of pronouns in the context of the WMT 2018 competition. We work with a test suite specifically designed to assess system quality in various fine-grained categories known to be problematic. The main evaluation scores come from a semi-automatic process, combining automatic reference matching with extensive manual annotation of uncertain cases. We find that current NMT systems are good at translating pronouns with intra-sentential reference, but the inter-sentential cases remain difficult. NMT systems are also good at the translation of event pronouns, unlike systems from the phrase-based SMT paradigm. No single system performs best at translating all types of anaphoric pronouns, suggesting unexplained random effects influencing the translation of pronouns with NMT.

1 Introduction

Data-driven machine translation (MT) systems are very good at making translation choices based on the words in the immediate neighbourhood of the word currently being generated, but aspects of translation that require keeping track of long-distance dependencies continue to pose problems. Linguistically, long-distance dependencies often arise from discourse-level phenomena such as pronominal reference, lexical cohesion, text structure, etc. Initially largely ignored, such problems have attracted increasing attention in the statistical MT (SMT) community in recent years (Hardmeier, 2012; Sim Smith, 2017). One important problem that has proved to be surprisingly difficult despite extensive research is the translation of pronouns (Hardmeier et al., 2015; Guillou et al., 2016; Loáiciga et al., 2017).

Since the invention of the BLEU score (Papineni et al., 2002), the MT community has measured progress to a large extent with the help of summary scores that are easy to compute, but strongly affected by the corpus-level frequency of certain phenomena, and that tend to neglect specific linguistic relations and problems that occur infrequently. The advent of neural MT (NMT) with its improved capacity for modeling more complex relationships between linguistic elements has brought an increased interest in linguistic problems perceived as difficult, which are often not captured well by metrics like BLEU. It has been suggested that test suites composed of difficult cases could provide more relevant insights into the performance of MT systems than corpus-level summary scores (Hardmeier, 2015). In this paper, we present a semi-automatic evaluation of the systems participating in the English–German news translation track of the MT shared task at the WMT 2018 conference.

The analysis was carried out with the help of an English–German adaptation of the PROTEST test suite for pronoun translation (Guillou and Hardmeier, 2016). The test suite allows us to perform a fine-grained evaluation for different types of pronouns. Whilst the translation of event pronouns, which caused serious problems in earlier evaluations of SMT systems (Hardmeier et al., 2015; Hardmeier and Guillou, 2018), seems to be handled fairly well by modern NMT systems, we find that translating anaphoric pronouns is still difficult, especially (but not only) if the pronoun has an antecedent in a different sentence. Our results also confirm earlier findings that suggested the need for a careful evaluation that is sensitive to specific linguistic problems. Whilst BLEU scores as a measure of general translation quality are strongly correlated with pronoun correctness, there are significant outliers that would be missed by an evaluation focusing on BLEU only. Moreover, evaluating pro-

[*]All authors contributed equally.

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 570–577

https://doi.org/10.18653/v1/W18-64062

noun translations by comparison with a reference translation is not reliable for all types of pronouns (Guillou and Hardmeier, 2018). This fact limits the usefulness of automatic pronoun evaluation metrics such as APT (Miculicich Werlen and Popescu-Belis, 2017) and affects the semi-automatic evaluation of our test suite as well.

2 Related Work

Research on pronoun translation was boosted by three past shared tasks (Hardmeier et al., 2015; Guillou et al., 2016; Loáiciga et al., 2017). They focused on English, French, German and Spanish in different directions. To avoid the effort and cost of manual evaluation, the tasks were designed and evaluated as classification rather than MT tasks, except for the first year, which featured both MT and classification tasks. At the time of the first of these shared tasks, phrase-based SMT systems were still competitive and the winning system was a strong n-gram language model (not involving any translation) trained as a baseline. By the time of the last pronoun focused shared task, however, an NMT system with no explicit knowledge about pronouns ranked first (Jean et al., 2017).

Automatic metrics computed by matching the candidate and reference translations offer little explanation of the causes for error. Additionally, the neural architectures of current end-to-end systems make it difficult to find out where exactly a translation went wrong by inspection. Test suites ease the evaluation process in general, since they allow us to simultaneously measure quantitative performance and diagnose qualitative shortcomings with regard to the targeted set of problems.

Test suites assessing NMT have focused on contrastive pairs or sets of sentences automatically generated. These include Burlot and Yvon (2017), for the evaluation of morphology in the English-to-Latvian and to-Czech language pairs; Sennrich (2017), who evaluates noun phrase and subject-verb agreement, particle verbs, polarity, and transliteration; and Rios Gonzales et al. (2017) whose work concentrates on word sense disambiguation for the German-to-English and German-to-French pairs. The test suite used in our work is based on the PROTEST test suite, which was originally created for English–French by Guillou and Hardmeier (2016). Closest to our work is the test suite of English-to-French anaphoric pronouns and coherence and cohesion by Bawden et al. (2018). Their test suite includes 50 examples of contrastive pairs of sentences, which are manually created and targeted towards object pronouns.

3 Test Suite Construction

The data for our test suite was taken from the ParCorFull corpus (Lapshinova-Koltunski et al., 2018), a German-English parallel corpus manually annotated for co-reference. Although the corpus is designed for nominal co-reference, it includes annotations of two types of antecedents: entities and events. Entities can be either pronouns or noun phrases, whereas events can be verb phrases, clauses, or a set of clauses.

ParCorFull includes texts from TED talks transcripts and newswire data. Specifically, it includes the datasets used in the ParCor corpus (Guillou et al., 2014), the DiscoMT workshop (Hardmeier et al., 2016), and the test sets from the WMT 2017 shared task (Bojar et al., 2017).

We constructed a test suite of 200 pronoun translation examples for English–German with a focus on the ambiguous English pronouns *it* and *they* and the aim of providing a set of examples that represents the different problems machine translation researchers should consider. We extracted the examples from the TED talks section of ParCorFull.

The selection is based on a two-level hierarchy which considers pronoun *function* at the top level, followed by other pronoun attributes at the more granular lower level (for anaphoric pronouns only).

The English pronoun *they* functions as an anaphoric pronoun, whereas *it* can function as either an anaphoric (1), pleonastic (2), or event reference[1] pronoun (3), with each function requiring the use of different pronouns in German.

(1) a. The infectious disease that's killed more humans than any other is malaria. **It**'s carried in the bites of infected mosquitos.
    b. Jene Krankheit, die mehr Leute als jede andere umgebracht hat, ist Malaria gewesen. **Sie** wird über die Stiche von infizierten Moskitos übertragen.

(2) a. And **it** seemed to me that there were three levels of acceptance that needed to take place.
    b. Und **es** schien, dass es drei Stufen der Akzeptanz gibt, die alle zum Tragen kom-

[1] *Event* reference is more commonly known as abstract anaphora or discourse deixis.

men mussten.

(3) a. But I think if we lost everyone with Down syndrome, **it** would be a catastrophic loss.
b. Aber, wenn wir alle Menschen mit Down-Syndrom verlören, wäre **das** ein katastrophaler Verlust.

At the more granular lower level, anaphoric pronouns are subdivided according to the following attributes: whether the pronoun appears in the same sentence as its antecedent (intra-sentential) or a different sentence (inter-sentential), the antecedent is a group noun, the pronoun is in subject or non-subject position (*it* only), or an instance of *they* is used as a singular pronoun (for example, to refer to a person of unknown gender). An overview of the resulting categories is provided in Table 2.

The distribution of test suite examples over the pronoun categories in the hierarchy can be found in the first row of Table 3. The number of examples assigned to each category reflects a) the functional ambiguity of the pronoun *it*, b) the number of different translation options possible in German, and c) the number of pronouns in the corpus that belong to the category (for example, there are very few instances of *singular they* available). Within each category, we aim to create a balance in terms of the expected pronoun translation token. We achieve this by considering the translation of the set of possible candidates in the reference translation.

4 Evaluation Results

The evaluation included 10 systems submitted to the English–German sub-task of the WMT 2018 competition and 6 anonymized online translation systems. Among the WMT submissions, all of the systems are neural models, with the Transformer (Vaswani et al., 2017) being a popular architecture choice. Implementation details can be found in the system description papers published at WMT 2018.

4.1 Automatic Evaluation

We provide scores from two different automatic evaluation metrics for all systems in our dataset (see Table 1 and Figure 1). To give a general impression of the translation quality achieved by the various systems, we include the BLEU scores on the TED talks from which the test suite is derived. These scores differ from the BLEU scores of the official WMT evaluation because they are computed on a different test set, containing texts from a different domain. For a more pronoun-specific evaluation, we also compute APT scores (Miculicich Werlen and Popescu-Belis, 2017).[2] For better comparability, the set of pronouns evaluated by APT was restricted to the 200 items included in the test suite. Following the recommendations of Guillou and Hardmeier (2018), we did not define any "equivalent" pronouns in the APT metric, but counted exact matches only.

System	BLEU	APT
Microsoft-Marian	32.6 3	66.0 7
NTT	31.8 7	**70.0** 1
UCAM	32.3 5	69.0 2
uedin	30.7 9	68.0 4
MMT-prod	**33.2** 1	65.0 8
KIT	31.6 8	68.5 3
online-Z	32.5 4	66.5 6
online-B	32.7 2	62.5 10
online-Y	31.9 6	68.0 5
JHU	28.8 10	62.0 12
online-F	18.8 14	60.5 13
LMU-nmt	28.5 11	63.0 9
online-A	27.4 12	62.0 11
online-G	22.3 13	59.5 14
RWTH-UNS	13.7 15	54.5 15
LMU-uns	10.5 16	–

Table 1: Automatic evaluation results.

A regression fit between the BLEU scores obtained and the number of examples annotated as correct by each system indicates a strong correlation between the two (Figure 2; $r = 0.912$, $N = 16$, $p < 0.001$), as does a similar analysis for the APT score ($r = 0.887$, $N = 15$, $p < 0.001$). These results, however, should be taken with a grain of salt, as we argue further in Section 5.

4.2 Semi-automatic Evaluation

The semi-automatic evaluation method is a two-pass procedure. It is motivated by the observation that automatic reference-based methods can identify correct examples with relatively high precision, but low recall (Guillou and Hardmeier, 2018). The evaluation procedure relies on word alignments, which were generated automatically by running Giza++ (Och and Ney, 2003) in both directions with grow-diag-final symmetrization (Koehn et al., 2005). The word alignments for the examples in the reference translation were corrected manually.

In the first step, the candidate translations are matched against the reference translation to ap-

[2]The APT score could not be computed for the LMU-uns system because the scorer cannot handle completely untranslated sentences, which occur occasionally in the output of that system.

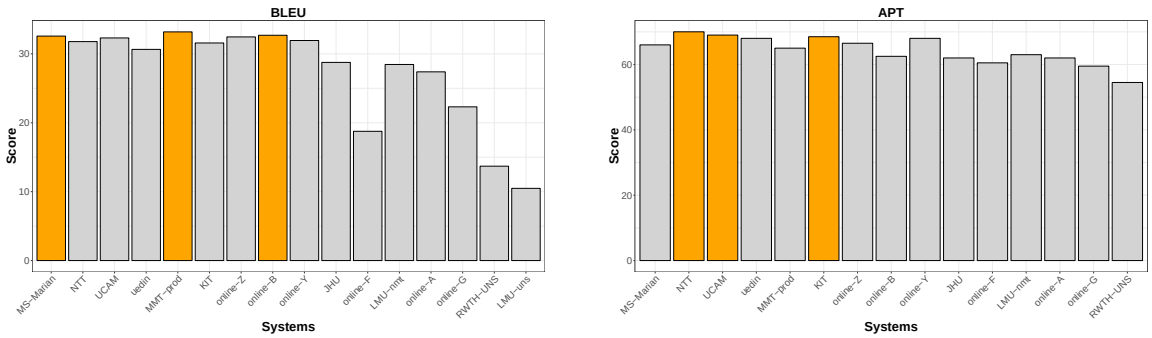

Figure 1: BLEU and APT scores. The three highest ranking systems are highlighted in orange.

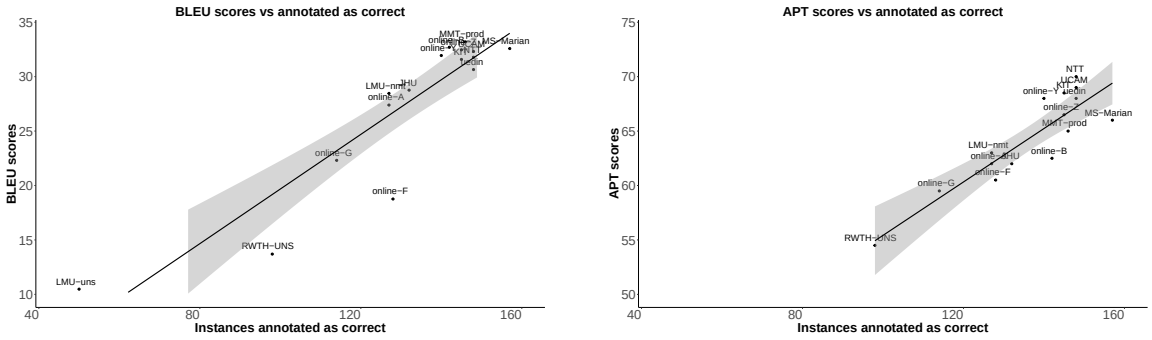

Figure 2: Correlation between the BLEU and APT scores and the number of instances annotated as correct. The gray zone indicates a 95% confidence interval.

prove examples that we can assume to be correct with reasonable confidence. Examples in the event and pleonastic categories can be approved based on a pronoun match alone; for the anaphoric categories, we also require matching antecedent translations. Two pronoun translations are considered to match if the sets of words aligned to the pronouns have at least one element in common after lowercasing. For antecedent translations, the word sequences aligned to the source antecedent must be completely equal for an automatic match. As a special exception, no automatic matches are generated for pronoun translations containing the word *sie* alone, so that the ambiguity between third-person plural *sie* and the pronoun of polite address *Sie* can be manually resolved.

In the second step, all examples not automatically approved are loaded into a graphical analysis tool specifically designed for the PROTEST test suite (Hardmeier and Guillou, 2016). The tool presents the annotator with the source pronoun, its translation by a given system, and the previous sentence for context. In the case of anaphoric pronouns, the context includes the sentence with the antecedent and one additional sentence. The examples were split randomly over four annotators. The annotators, who are translator trainees at Saarland University, are all native speakers of German with a good knowledge of English. To improve the quality of the annotations, the annotators had been trained beforehand on the output of a baseline NMT system.

Category	**–**	**+**	**total**	**correct**
Anaphoric				
intra-sent subj. *it*	5	39	44	88.6%
intra-sent non-subj. *it*	6	13	19	68.4%
inter-sent subj. *it*	13	16	29	55.2%
inter-sent non-subj. *it*	9	21	30	70.0%
intra-sent *they*	–	–	–	–
inter-sent *they*	–	–	–	–
singular *they*	–	–	–	–
group *it/they*	–	9	9	100.0%
Event reference *it*	14	68	82	82.9%
Pleonastic *it*	–	137	137	100.0%
Total	47	303	350	86.6%

Table 2: Human evaluation of automatically approved examples

In total, 3,200 pronoun examples from 16 systems were evaluated. 1,150 examples were approved automatically and 2,050 examples were referred for manual annotation. To verify the validity of the semi-automatic method, we also solicited manual annotations for a random sample of 350 examples that had been approved automatically.

The first step of our two-step procedure can only approve examples, it never rejects them automatically. As a consequence, our semi-automatic evaluation is *biased towards correctness* with respect to a fully manual evaluation. The scores presented in Table 3 will therefore tend to overestimate the actual system performance.

The results of the human annotation of the random sample of 320 examples automatically matched as correct are presented in Table 2. Consistently with similar results for French (Hardmeier and Guillou, 2018), 86.6% of the automatically approved examples were accepted as correct by the evaluators. However, we must highlight that the accuracy of the automatic evaluation varies substantially across categories. Whilst pronouns known to be pleonastic can be checked automatically with very good confidence, the automatic evaluation of anaphoric pronouns is much more difficult, with an evaluation accuracy as low as 55.2% in the inter-sentential subject *it* case. This reflects the general difficulty of automatic pronoun evaluation (Guillou and Hardmeier, 2018) and reinforces the positive bias discussed in the previous paragraph for these categories in particular.

The results of the semi-automatic evaluation are displayed in Table 3. For the counts in this table, we used *manual* annotations wherever possible. Automatic annotations were used only for those examples that had not been annotated manually.

The best result was obtained by the Microsoft-Marian system, which translated 157 out of 200 pronouns correctly. It is followed by a group of 5 shared task submissions that achieved scores between 145 and 148. Three of the online systems also reached scores over 140. The remaining shared task submissions are JHU with a score of 132 and LMU-nmt with a score of 127. Unsurprisingly, the unsupervised submissions are ranked last.

5 Discussion

Generally speaking, a high BLEU score indicates good translation quality and vice versa. The APT score has been shown to capture good pronoun translations with reasonable precision, if unsatisfactory recall (Guillou and Hardmeier, 2018), but it is also trivially correlated with our test suite score to some extent because the automatic part of our semi-automatic evaluation identifies good translation with a mechanism that is very similar to that of APT. In the right half of Figure 2, we observe that the APT score introduces spurious differences between systems reaching exactly the same number of correctly translated items (NTT, UCAM, uedin) and fails to reward correct pronoun translations in some of the systems (Microsoft-Marian, online-B). As a result, the score can serve as an indicator, but not as a reliable replacement of a manual or semi-automatic evaluation.

Moreover, the small size of the test suite and the differences between the system architectures must be kept in mind. Considering these two factors, a larger threshold in any of the two scores is needed to claim that one system is actually better than another (Berg-Kirkpatrick et al., 2012). This caveat appears to be confirmed by the two outliers seen in the left part of Figure 2. Interestingly, the online-F system achieves many good pronoun translations despite a low BLEU score. The RWTH-uns system is also much better on correct pronouns than LMU-uns (the other unsupervised system) than the difference in BLEU scores would suggest.

The results of manual evaluation vary significantly by category. In the anaphoric *it* categories, it is evident that intra-sentential anaphora is easier to handle than inter-sentential anaphora. In the intra-sentential case, the best systems produce correct translations for 70–80% of the examples, which is a fair result, but indicates that the problem is not completely solved yet. In the inter-sentential *it* categories, the average performance is below 50% despite the positive bias of our evaluation method, and even the best-performing systems are not much better. It is worth noting that no single system performs best over all anaphoric categories, which suggests that the top scores achieved for this part of the test suite could be random strokes of luck. The results for pronouns in subject and non-subject positions are not very different. This contrasts with the results of Hardmeier and Guillou (2018) for English–French, where non-subject pronouns were found to be substantially harder to translate. It might be due to the fact that the direct object forms of French personal pronouns coincide with those of the definite article, a problem that does not apply to German.

The plural cases of *they* do not cause any serious problems, at least for the stronger systems, since *they* can usually be translated straightforwardly using the German pronoun *sie*. The errors occurring in these categories are often due to confusion with the pronoun of polite address *Sie* (“you”). When

	Pronouns											Antecedents
	anaphoric								event	pleonastic		
	it				they			it/they	it	it		
	intra		inter		intra	inter	sing.	group				
	subj.	non-subj.	subj.	non-subj.							**Total**	
Examples	*25*	*25*	*25*	*25*	*10*	*10*	*5*	*15*	*30*	*30*	***200***	*140*
Microsoft-Marian	18	20	12	15	9	10	2	13	29	29	**157**	132
NTT	16	18	14	16	10	10	1	8	26	29	**148**	135
UCAM	19	20	13	11	10	10	2	11	22	30	**148**	134
uedin	19	19	10	11	10	10	–	11	29	29	**148**	132
MMT-prod	20	19	11	15	10	8	–	9	25	29	**146**	137
KIT	19	18	15	11	9	9	1	6	27	30	**145**	126
online-Z	21	18	10	10	10	10	2	11	24	29	**145**	132
online-B	20	15	12	12	8	10	–	8	27	30	**142**	128
online-Y	18	17	11	12	10	9	1	8	24	30	**140**	136
JHU	12	17	8	11	8	10	3	10	24	29	**132**	119
online-F	13	16	10	11	10	10	2	7	21	28	**128**	115
LMU-nmt	10	9	10	13	7	10	1	9	28	30	**127**	125
online-A	11	9	12	16	5	10	2	5	27	30	**127**	130
online-G	10	6	15	11	2	8	2	7	23	30	**114**	119
RWTH-uns	9	5	9	8	3	8	1	7	19	29	**98**	99
LMU-uns	4	2	2	2	4	8	–	5	15	8	**50**	87
Average												
count	14.9	14.3	10.9	11.6	7.8	9.4	1.3	8.4	24.4	28.0	**130.9**	124.1
percentage	59.8	57.0	43.5	46.3	78.1	93.8	25.0	56.3	81.3	93.5	**65.4**	88.6

Table 3: Pronoun and antecedent translations marked as correct, per system

they has a singular antecedent or refers to a group, however, it is mistranslated much more frequently.

The only system that has noticeable problems with pleonastic *it* is the unsupervised LMU-uns submission. Translating event *it* seems to be more difficult, but many systems still achieve close to perfect results in this category. Similarly to the results of Hardmeier and Guillou (2018) for English–French, this suggests that NMT systems are quite good at identifying pronouns with event reference and producing appropriate translations for them.

6 Conclusions

We have presented a detailed analysis of 16 NMT systems, assessing their performance in the translation of pronouns using a semi-automatic evaluation based on a balanced test suite. The results reinforce the idea that automatic evaluation scores are correlated with manual evaluation results, but they also confirm that automatic evaluation can provide a misleading picture of the behavior of some systems. The evaluation has also reinforced that special attention should be paid to the problematic cases that are only identifiable through the careful balance of categories achieved in the test suite design. This balanced design has also made us aware of the progress made by NMT in the modeling of context for the translation of pleonastic, event and intra-sentential anaphoric pronouns. Pleonastic pronouns are handled almost perfectly by most systems, so we suggest that future evaluations emphasize the more challenging cases. Anaphoric pronouns depending on the inter-sentential context remain a significant challenge. They present an ideal test case for the development of context-aware NMT systems. Research in that direction has recently gained some traction (Tiedemann and Scherrer, 2017; Wang et al., 2017; Tu et al., 2018) and has claimed promising results specifically for pronoun translation (Voita et al., 2018). It remains to be seen whether the development of such methods will lead to a breakthrough in the translation of inter-sentential anaphoric pronouns in the near future.

Acknowledgements

The work carried out at Uppsala University was supported by the Swedish Research Council under grants 2012-916 (to Jörg Tiedemann) and 2017-930 (to Christian Hardmeier). We thank Jörg Tiedemann for helping us fund this effort. The work carried out at The University of Edinburgh was funded by the ERC H2020 Advanced Fellowship GA 742137 SEMANTAX and a grant from The

University of Edinburgh and Huawei Technologies. The manual test suite evaluation was funded by the European Association for Machine Translation. We thank our annotators Daria Hert, Georg Seiler, Alexander Schütz and Peter Schneider for their valuable work.

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Eleventh Language Resources and Evaluation Conference*, LREC 2016, pages 636–643, Portorož, Slovenia.

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP, Brussels, Belgium. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, pages 525–542, Berlin, Germany. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 2014. ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC 2014, pages 3191–3198, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours*, 11.

Christian Hardmeier. 2015. On statistical machine translation and translation theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 168–172, Lisbon (Portugal). Association for Computational Linguistics.

Christian Hardmeier and Liane Guillou. 2016. A graphical pronoun analysis tool for the protest pronoun evaluation test suite. *Baltic Journal of Modern Computing*, 4(2):318–330.

Christian Hardmeier and Liane Guillou. 2018. Pronoun translation in English–French machine translation: An analysis of error types. *ArXiv e-prints*, 1808.10196.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

Christian Hardmeier, Jörg Tiedemann, Preslav Nakov, Sara Stymne, and Yannick Versely. 2016. DiscoMT 2015 shared task on pronoun translation. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. `http://hdl.handle.net/11372/LRT-1611`.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Neural machine translation for cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 54–57, Copenhagen, Denmark. Association for Computational Linguistics.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of 11th Language Resources and Evaluation Conference*, pages 423–428, Miyazaki, Japan. European Language Resources Association (ELRA).

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 DiscoMT shared task on cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 1–16, Copenhagen, Denmark. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL 2002, pages 311–318, Philadelphia. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? Assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Karin Sim Smith. 2017. On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.