# Large-scale Machine Translation Evaluation of the iADAATPA Project

**Sheila Castilho,* Natália Resende,* Federico Gaspari,* Andy Way***
**Tony O'Dowd,† Marek Mazur,† Manuel Herranz,§ Alex Helle,§**
**Gema Ramírez-Sánchez,‡ Víctor Sánchez-Cartagena,‡ Mārcis Pinnis,** Valters Šics****
*ADAPT Centre, Dublin City University firstname.lastname@adaptcentre.ie
†KantanMT tonyod/marekm@kantanmt.com
§Pangeanic m.herranz/a.helle@pangeanic.com
‡Prompsit gramirez/vmsanchez@prompsit.com
** Tilde marcis.pinnis/valters.sics@tilde.lv

## Abstract

This paper reports the results of an in-depth evaluation of 34 state-of-the-art domain-adapted machine translation (MT) systems that were built by four leading MT companies as part of the EU-funded iADAATPA project. These systems support a wide variety of languages for several domains. The evaluation combined automatic metrics and human methods, namely assessments of adequacy, fluency, and comparative ranking. The paper also discusses the most effective techniques to build domain-adapted MT systems for the relevant language combinations and domains.

## 1 Introduction

The evaluation reported in this paper was conducted as part of the EU-funded iADAATPA (intelligent, Automatic Domain-Adapted Automated Translation for Public Administrations) project that ended in February 2019.[1] The evaluation was performed by the ADAPT Centre at Dublin City University (DCU) on 34 state-of-the-art domain-adapted machine translation (MT) systems built by four leading MT companies KantanMT, Pangeanic, Prompsit and Tilde. These MT engines supported a wide range of language pairs, including under-resourced ones, for several domains.

The main objective of the project was to lower language barriers with a view to promoting truly multilingual services across EU Member States. To this end, an innovative platform (MTHub) was developed to offer state-of-the-art domain-adapted MT engines to public administrations (PAs) in addition to the EU's own eTranslation service.[2] In this context, the technical partners of the project built MT engines for the language pairs and in the specific domains that were indicated as priorities by the PAs in the respective countries.

The rest of the paper is structured as follows. The four MT companies involved in this study are presented in Sections 2 (KantanMT), 3 (Pangeanic), 4 (Prompsit), and 5 (Tilde), with a description of the systems that they developed, including the data that they used and how they customized their engines. Section 6 outlines the protocol that was followed for this large-scale automatic and human evaluation. Section 7 reports the key results of the evaluation, and Section 8 concludes with a summary of the most important lessons learned and possibilities for further work in this area.

## 2 KantanMT

KantanMT offers a cloud-based MT platform that enables users to develop and manage customized MT engines. The technologies offered enable users to build MT engines in over 750 language combinations integrated into the user's localisation workflows and web applications.

### 2.1 KantanMT's MT systems

**Language pairs and domains** KantanMT's PA partner was DCU, whose website encompasses more than 120 sub-sites providing informational content for students, lecturers and visitors to the DCU Campus. Due to the amount of content on



[1] http://iadaatpa.com/

[2] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

| Language pair | # segments |
|---|---|
| | train |
| English→Bulgarian | 8.3M |
| English→Dutch | 11.8M |
| English→French | 13.6M |
| English→German | 41.2M |
| English→Irish | 1.8M |
| English→Italian | 5.2M |
| English→Polish | 10.7M |
| English→Portuguese | 8.1M |
| English→Romanian | 8.9M |
| English→Spanish | 18.5M |

**Table 1:** Data used to train KantanMT NMT systems

the University's website, a small number of areas were prioritized: News and Events, President Office, School of Applied Languages and Intercultural Studies, and Fiontar – Irish Language Research. The language pairs consisted of English as the source for all the neural MT (NMT) engines into Bulgarian, Dutch, French, German, Irish, Italian, Polish, Portuguese, Romanian, and Spanish.

### 2.2 Data Acquisition

The data used in the customization of KantanMT's engines was selected from publicly available sources, such as the DGT (European Commission's Directorate-General for Translation), EMEA (European Medicines Agency), ECB (European Central Bank) and EuroParl (Koehn, 2005).[3] Table 1 shows the training data for KantanMT's NMT systems.

### 2.3 Engine Customization

All initial NMT engines were developed using the Torch implementation of the OpenNMT framework.[4] The development test reference set, used to generate automated scores and to establish a performance baseline for each engine, consisted of 500 segments chosen at random from the live DCU website. Both recurrent and transformer neural models were trained. The model with the best overall automated scores was then selected as the final release candidate. (For the purposes of engine selection, F-Measure, TER (Snover et al., 2006), BLEU (Papineni et al., 2002), and Perplexity were used as automated scores.)

[3]https://www.statmt.org/europarl/
[4]http://www.opennmt.net/

| Language pair | # segments | |
|---|---|---|
| | init | train |
| Spanish→Catalan | 30k | 13.6M |
| Spanish→English | 30k | 14.6M |
| Spanish→French | 30k | 14.6M |
| Spanish→German | 30k | 14.5M |
| Spanish→Italian | 30k | 14.5M |
| Spanish→Portuguese | 30k | 14.6M |
| Spanish→Russian | 30k | 13.8M |

**Table 2:** Data used to train Pangeanic NMT systems

## 3 Pangeanic

Pangeanic (Yuste et al., 2010) is a Language Service Provider (LSP) specialised in Natural Language Processing and MT. It provides solutions to cognitive companies, institutions, translation professionals, and corporations.

### 3.1 Pangeanic's MT systems

**Language pairs and domains** Pangeanic's use-cases were for two Spanish PAs: (1) Generalitat Valenciana (regional administration) translating from Spanish into and out of English, French, Catalan/Valencian, German, Italian, Russian; and (2) Segittur (tourism administration) translating from Spanish into and out of English, French, German, Italian, Portuguese. For this purpose, NMT systems were built.

**Data acquisition** For Spanish→Russian there was no available in-domain data. Therefore, two translators were contracted as part of the project to create 30k segments of in-domain data, translating PAs' websites. They also cleaned UN material and post-edited general-domain data that was previously filtered as in-domain following the "Invitation Model" (Hoang and Sima'an, 2014). For the other language pairs, the input material was also 30k post-edited segments. The main part of the training corpora (approximately 75%) came from Pangeanic's own repository, harvested through web crawling including OpenSubtitles (Tiedemann, 2012). The rest of the corpus was automatically validated synthetic material using general data from Leipzig (Goldhahn et al., 2012). Table 2 shows the size of the in-domain (manual or provided by the PA) and generic training data set.

**Engine customization** The data was cleaned using the Bicleaner tool (Sánchez-Cartagena et al., 2018). Moreover, embeddings for case infor-

| Use-case | Language pair | # segments | |
|---|---|---|---|
| | | init | train |
| Gazette | Spanish→English | 0 | 34.2M |
| Gazette | Spanish→Basque | 820k | 820k |
| R&D | English→Spanish | 0 | 36.3M |
| R&D | Basque→Spanish | 0 | 4.6M |

**Table 3:** Data used to train Prompsit NMT systems

mation and byte pair encoding tokenization were added. The models were trained with multi-domain data and we improved performance following a domain-mixing approach (Britz et al., 2017). The domain information was indicated using special tokens for each target sequence. The domain prediction was based only on the source as the extra token was added at target-side and there was no need for *a priori* domain information. This approach allowed the model to improve the quality for each domain.

## 4 Prompsit

Prompsit is a language technology (LT) provider with a strong focus on tailored MT services involving data curation, training and development of other multilingual applications.

### 4.1 Prompsit's MT systems

**Language pairs and domains** Prompsit partnered with SESIAD, the Spanish State Secretary for Information Society and Digital Agenda, and built eight MT systems for two use-cases: (1) translation of the Spanish Official Gazette from Spanish into Catalan, Galician, Basque and English and (2) translation of R&D content for monitoring purposes from Catalan, Galician, Basque, and English into Spanish. Rule-based MT (RBMT) was used for combinations involving Catalan and Galician (mainly due to the lack of relevant corpora) and NMT was used for the rest.

**Data acquisition** For the RBMT systems, monolingual and bilingual data was crawled from different websites. For the NMT systems, data was compiled by means of web crawling, back- and forward-translation of monolingual corpora, and cross-entropy data selection. Table 3 presents the amount of in-domain parallel segments initially available and finally used to train NMT systems.

**Engine customization** RBMT systems based on Apertium (Forcada et al., 2011) were customized by extracting candidates for new monolingual and bilingual dictionary entries from a word-aligned parallel corpus generated with ruLearn (Sánchez-Cartagena et al., 2016). For NMT systems, based on OpenNMT, automatic segmentation of long sentences and linguistically informed word segmentation for Basque (Sánchez-Cartagena, 2018) were added to the corpus pre-processing pipeline. Moreover, to ensure translation consistency, carefully designed terminology to restrict translation hypotheses and named entity recognition to control the translation of proper names, places, etc. was added. Finally, mixed fine-tuning (Chu et al., 2017) was applied to some systems to balance the weight of the different sources of training data.

## 5 Tilde

Tilde is an LSP and LT developer offering customized MT system development, as well as a wide range of other cloud-based and stand-alone LT tools and services for terminology management, spelling and grammar checking, speech recognition and synthesis, personalised virtual assistants, and other applications. It provides on-premise and cloud-based LT solutions to public and private organisations as well as LT productivity tools to individual users.

### 5.1 Tilde's MT systems

**Language pairs and domains** The use-cases for Tilde cover political news for English into Bulgarian and Estonian, general news for English into Latvian and legislation (legal acts and legislative news) for English into Lithuanian and Lithuanian into Russian. The Lithuanian language use-cases are intended for the Seimas of the Republic of Lithuania (the Parliament of Lithuania).

**Data acquisition** All NMT systems were trained using a combination of broad domain corpora and synthetic in-domain corpora (i.e. back-translated monolingual corpora). The in-domain corpora were acquired by crawling relevant web domains containing in-domain data as well as by acquiring translation memories from the partner PA. All parallel corpora were normalized, cleaned from noise, and pre-processed using the methodology by Pinnis et al. (2017). The training data statistics for the NMT systems are provided in Table 4.

**Engine customization** At first, initial NMT systems were trained using Nematus (Sennrich et

| Use-case | Lang. pair | # segments | |
|---|---|---|---|
| | | Init | Domain |
| General news | Eng.→Lat. | 15.8M | 11.6M |
| | Lat.→Eng. | 15.8M | 11.1M |
| Political news | Eng.→Est. | 18.9M | 1.7M |
| | Est.→Eng. | 18.9M | 0.7M |
| | Eng.→Bul. | 6.2M | 6.2M |
| | Bul.→Eng. | 6.2M | 6.1M |
| Law | Eng.→Lit. | 10.2M | 0.5M |
| | Lit.→Eng. | 10.2M | 10.1M |
| | Lit.→Rus. | 2.7M | 2.6M |
| | Rus.→Lit. | 2.7M | 2.6M |

**Table 4:** Data used to train Tilde NMT systems

al., 2017) with the multiplicative long short-term memory unit implementation by Pinnis et al. (2017). Then, monolingual in-domain data were back-translated (Poncelas et al., 2018). For systems for which the in-domain data amounted to the same amount as the initial training data, the back-translated synthetic parallel corpora were added to the initial training data and final (domain-specific) systems were trained from scratch. For the remaining systems (English-Estonian and English-Lithuanian), domain adaptation of the initial models was performed using continued training.

## 6 Evaluating iADAATPA's MT Systems

The evaluation of all iADAATPA's MT systems was carried out following current MT assessment practices (see Castilho et al. (2018)) with a combination of automatic evaluation metrics (AEMs) – including BLEU, METEOR (Banerjee and Lavie, 2005), TER and chrF (Popović, 2015) – and human evaluation, consisting of assessing fluency, adequacy and ranking against a baseline. The *Adequacy* rating was based on the statement "The translated sentence conveys the meaning of the original...", which was to be completed with a 3-point Likert scale (1-Poorly, 2-Fairly, 3-Well). The *Fluency* rating was based on the statement "The translated sentence is grammatically...", which was to be completed with a 3-point Likert scale (1-Incomprehensible, 2-Fair, 3-Flawless). The *Ranking* assessment was based on asking the translators to rate the translations from best to worst. Ties were allowed for both "equally well translated" or "equally badly translated".

The baseline MT system selected to be compared against the partners' engines for both automatic and human evaluation was Google Translate (GNMT).[5] However, for KantanMT's systems, the baseline chosen for the human evaluation was the human reference translations; this choice was made as the systems were not in their final version by the time they were delivered for evaluation, so the partner was keen to know initially how their systems performed against a gold standard in order to subsequently improve them.

### 6.1 Test Sets

The test sets consisted of 500 randomized sentences and were provided by the MT partners. A portion of these data sets was used to compute inter-annotator agreement (IAA, see Section 7.2.1). The partners also provided the reference translations for the source texts, which were translated professionally.

### 6.2 Translators

Each system was evaluated by two professional translators, who did not know whether the translations were from the partner's MT system or the baseline. Guidelines on how to use the evaluation tools and how to assess the translations were provided. IAA was computed on sets of 100 sentences; however, blank data points (skipped evaluations or bugged data points) were removed from the raw data set, which led to a variance in the total number of sentences.

### 6.3 Tool

The tool used to assess fluency, adequacy and ranking was KantanMT's LQR,[6] a cloud-based platform which facilitates the interaction with translators since they are not required to download any software.

## 7 Results

### 7.1 Automatic Evaluation Metrics

Due to space constraints, here we present average scores of the MT engines' AEM results grouped by partner (Engine) against average scores of GNMT (Baseline), pointing out particularly interesting aspects.

KantanMT's MT systems (Fig.1) score higher than GNMT in the majority of the cases, with the exception of the English-Italian system which does not outperform the GNMT system.

[5] https://translate.google.com/

[6] https://www.kantanmt.com/overview-kantanlqr.php

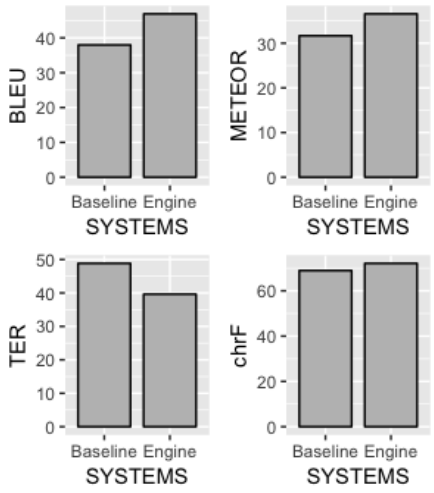


**Figure 1:** Automatic metrics - KantanMT

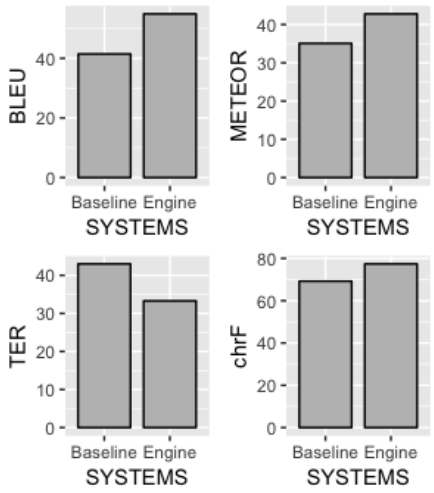


**Figure 2:** Automatic metrics - Pangeanic

Pangeanic's MT systems (Fig.2) score higher than GNMT in almost all cases. The Spanish-English MT system is the only one that does not outperform the baseline by a statistically significant margin, possibly as a result of the Spanish PA's content being overly generic and thus competing on a general basis against GNMT as opposed to a customized engine.

Prompsit's MT systems (Fig.3) score higher than GNMT for the majority of the cases.

Tilde's MT systems (Fig.4) score higher than GNMT in most cases, with Latvian-English and English-Latvian being the only the engines that do not outperform the baseline (by a statistical significant amount for Latvian↔English).

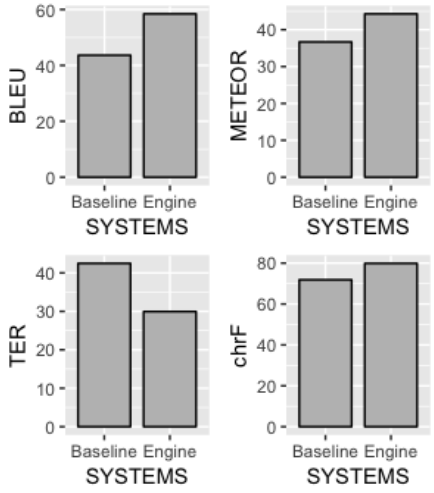


**Figure 3:** Automatic metrics - Prompsit

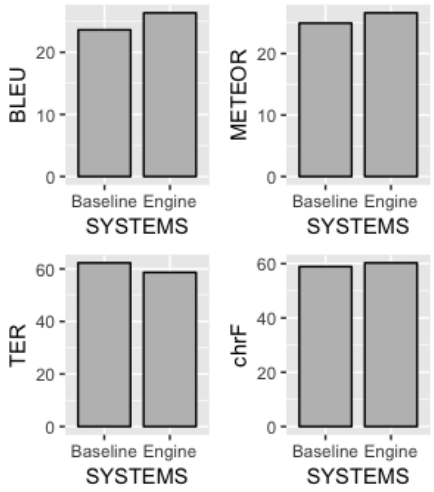


**Figure 4:** Automatic metrics - Tilde

### 7.2 Human Evaluation

#### 7.2.1 Inter-Annotator Agreement

Overall, an average kappa coefficient between 0.21 and 0.40 (moderate) and between 0.40 and 0.60 (fair) was observed for fluency and adequacy for both weighted and non-weighted kappa for all partners' engines and baseline. Poor agreement (k=0.0-0.20) was observed only for non-weighted kappa for fluency ratings of KantanMT's baseline, adequacy ratings of Pangeanic's engines and baseline, and adequacy ratings of Prompsit's engines and baselines.

#### 7.2.2 Fluency, Adequacy and Ranking

The results for fluency, adequacy and ranking show that the iADAATPA partners' MT systems systematically outperformed GNMT. These results mean that our partners' systems were considered

better than GNMT most of the time and that their output was deemed to be grammatically more fluent and more adequate compared to the source sentences than GNMT's output. The only exception observed is for KantanMT's MT systems (Figure 5), which did not outperform the baseline; however, this was an expected result since the baseline for KantanMT's systems was the human reference translation. In the interest of conciseness, Figures 5, 6, 7 and 8 illustrate the average performance of all partners' MT systems combined (Engine), arranged by company, against the respective baselines.

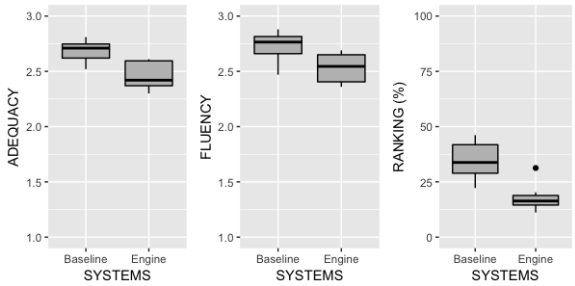


**Figure 5:** Human evaluation - KantanMT partner

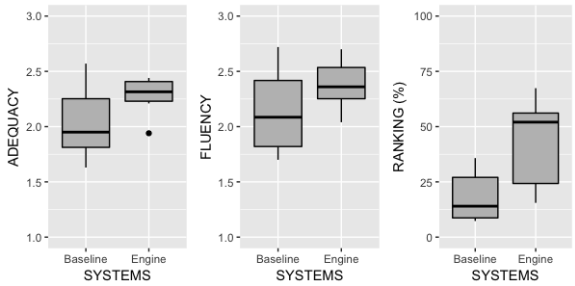


**Figure 6:** Human evaluation - Pangeanic partner

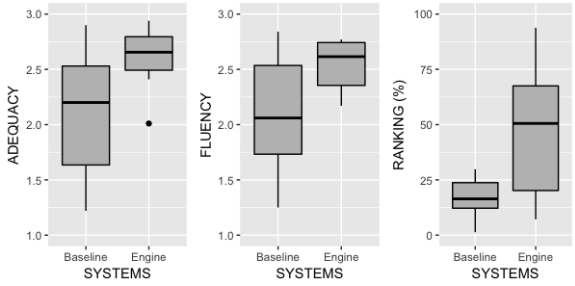


**Figure 7:** Human evaluation - Prompsit partner

## 8 Conclusion

The results of this comprehensive evaluation show that in general the MT systems developed within the iADAATPA project were competitive with the production systems, including for language pairs that lack extensive resources. In particular, the evaluation with the four standard AEMs consistently showed the partners' MT systems to have superior performance compared to the baseline engines. In addition, the human evaluation of fluency and adequacy as well as comparative ranking also yielded very positive results; with the exception of the MT systems developed by KantanMT, which were compared against the human reference baseline, all the other domain-adapted engines prevailed in the human evaluation, with a clear preference over the baseline in the comparative ranking.

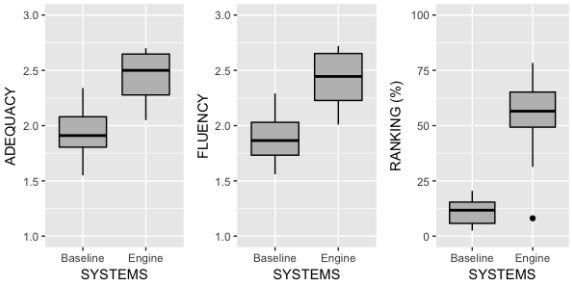


**Figure 8:** Human evaluation - Tilde partner

The evaluation presented here can be extended in several ways, e.g. including the results for updated versions of the MT systems covered in these experiments; during the iADAATPA project the systems were continuously improved with additional training data and more sophisticated techniques, to optimize their performance vis-à-vis the targeted use-cases indicated by the respective PAs. In addition, we intend to investigate the relationship between the additional development efforts and the improved performance, especially in terms of automatic metrics, as conducting additional human evaluation is unlikely, given that the project is now concluded.

**Acknowledgements** The work reported in this paper was conducted during the iADAATPA project, which was funded by INEA through grant N° 2016-EU-IA-0132 as part of the EU's CEF Telecom Programme. The ADAPT Centre for Digital Content Technology at Dublin City University is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

## References


Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, Ann Arbor, MI.

Britz, Denny, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine Translation Quality Assessment. In Castilho, Sheila, Joss Moorkens, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, pages 9–38. Springer.

Chu, Chenhui, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim ORegan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC12)*, pages 759–765, Istanbul, Turkey.

Hoang, Cuong and Khalil Sima'an. 2014. Latent domain translation models in mix-of-domains haystack. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 1928–1939, Dublin, Ireland.

Koehn, Philipp. 2005. Europarl: a parallel corpus for statistical machine translation. In *MT Summit X, Conference Proceedings: the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Papineni, Kishore, 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 on Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Pinnis, Mārcis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics. 2017. Tilde's Machine Translation Systems for WMT 2017. In *Proceedings of the Second Conference on Machine Translation, Vol. 2: Shared Task Papers*, pages 374–381, Copenhagen, Denmark.

Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In *Proceedings of The 21st Annual Conference of the European Association for Machine Translation*, pages 249–258, Alicante, Spain.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Sánchez-Cartagena, Víctor M., Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2016. Rulearn: an open-source toolkit for the automatic inference of shallow-transfer rules for machine translation. *The Prague Bulletin of Mathematical Linguistics*, 106(1):193–204.

Sánchez-Cartagena, Víctor M., Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Vol. 2: Shared Task Papers*, pages 955–962, Brussels, Belgium.

Sánchez-Cartagena, Vıctor M. 2018. Prompsits submission to the IWSLT 2018 low resource machine translation task. In *Proceedings of 15th International Workshop on Spoken Language Translation*, pages 95–103, Bruges, Belgium.

Sennrich, Rico, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Others. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the EACL*, pages 65–68, Valencia, Spain.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 2214–2218, Istanbul, Turkey.

Yuste, E., M. Herranz, A. Helle, and H. Suzuki. 2010. Pangeamt - putting open standards to work ... well. In *AMTA 2010: the Ninth conference of the Association for Machine Translation in the Americas*, Denver, CO. 8pp.