CEF Telecom

**Project title:** MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages

# Evaluation of data release 1 completed

## Milestone number: 7

Version 1.0

**Co-financed by the Connecting Europe Facility of the European Union**

Funded by the European Union's CEF Telecom programme
under the Grant Agreement No. 2020-EU-IA-0078

| | |
|---|---|
| **Project Acronym:** | MaCoCu |
| **Project Full Title:** | MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languag |
| **Year of the Call:** | 2020 |
| **Grant Number:** | 2020-EU-IA-0078 |
| **Project URL:** | https://macocu.eu |

| | |
|---|---|
| Document title: | Evaluation of data release 1 completed |
| Lead author: | Rik van Noord (Rijksuniversiteit Groningen) |
| Contributing authors: | Rik van Noord, Antonio Toral (Rijksuniversiteit Groningen)<br>Miquel Esplà-Gomis (Universitat d'Alacant)<br>Gema Ramírez-Sánchez (Prompsit)<br>Peter Rupnik, Taja Kuzman, Nikola Ljubešić (JSI) |
| Milestone number: | 7 |
| Activity associated to the milestone: | Evaluation |
| Dissemination level: | Public (PU) |
| Contractual Delivery Date: | May 31, 2022 |
| Actual Delivery Date | May 31, 2022 |
| Number of pages: | 15 |

## Document history

| Version | Date | Changes |
|---|---|---|
| 1.0 | May 31, 2022 | Original Submission |

## Abstract

This report provides a description of the evaluation process carried out to assess the usefulness of the first data release of the MaCoCu project. This data release contains monolingual data for Maltese, Croatian, Icelandic, Macedonian, Turkish, Slovene and Bulgarian, and parallel data for the same languages aligned with English. The corpora have been evaluated in a number of natural-language-processing tasks in order to compare them to the performance obtained with other state-of-the-art corpora and pretrained models. The results obtained confirm the usefulness of the corpora released both in a monolingual and a bilingual context of use in a number of ways: (1) training on monolingual MaCoCu data clearly improves performance of pretrained language models across a variety of tasks and languages; (2) adding parallel MaCoCu data to a baseline neural machine translation model improves performance for all evaluated languages and (3) a human evaluation confirms the findings of (2) in that professional translators generally prefer translations produced by a partially MaCoCu-trained model over a non-MaCoCu baseline.

# Contents

# 1 Executive summary

## 1.1 Introduction

This deliverable, submitted as Milestone 7: *Evaluation of data release 1*, corresponds to the verification means for Activity 7: *Evaluation* for the first year of the project. The main goal of Activity 7 is to evaluate the usefulness of the data produced through Activities 5: *Curation of bilingual data*, and 6: *Curation of monolingual data*, to produce the corpora that were released as part of the first data release of the project on April 30.

The report is divided into two main blocks: evaluation of monolingual data, and evaluation of parallel data. For parallel data, the evaluation carried out focuses mostly on the task of machine translation, and covers two scenarios: automatic evaluation and human evaluation of the resulting translation systems. For automatic evaluation, state-of-the-art models were trained combining different corpora, including or not the MaCoCu corpora released; then, these models were automatically evaluated on available test sets from the Flores data set [1], using automatic evaluation metrics such as COMET [2] and BLEU [3]. For human evaluation, a team of translators was asked to rank the translations produced by models trained with and without the MaCoCu corpora.

In the case of monolingual corpora, the data were used to train large general-purpose language models, in the same format as the popular models BERT [4] and RoBERTa [5]. These were then applied to several natural-language-processing tasks, such as part-of-speech tagging, named-entity recognition and plausible alternative classification.

## 1.2 Brief summary of the MaCoCu action

The objective of MaCoCu is to gather, clean and enrich monolingual and parallel corpora for several European languages with scarce resources. This will be achieved by crawling, cleaning and adding extra info to data in multiple languages.

Four European partners from academia and industry, all highly specialised, take part in this action. All of them play a crucial role in the development of the objectives of the action.

The four partners are:

- University of Alacant (UA): responsible for code and project management.
- Jožef Stefan Institute (JSI): responsible for data crawling and monolingual curation and enrichment.
- Rijksuniversiteit Groningen (RUG): responsible for DSI-specific content enrichment and evaluation.
- Prompsit Language Engineering, SL (Prompsit): responsible for bilingual data curation and dissemination and outreach.

Partners RUG, JSI and Prompsit were involved in Activity 7, on which this report is focused.

This report covers the first batch of languages in the project, which includes: Bulgarian, Icelandic, Macedonian, Maltese, Croatian, Slovene and Turkish.

# 2 Evaluation of monolingual corpora

## 2.1 Automatic evaluation of monolingual corpora

This section focuses on the evaluation of the monolingual corpora that were part of the first MaCoCu release. We evaluate the corpora by training general purpose language models (LMs). Pre-trained LMs are state-of-the-art in virtually all NLP tasks, and we see the opportunity to build better such LMs than currently exist for some of MaCoCu's languages. These thus could be expected to be adopted by the NLP community. Similar models for, among others, Dutch [6] and French [7] are currently very popular in the community.

We train such pretrained LMs for five languages: Bulgarian/Macedonian, Icelandic, Maltese and Turkish. We train a single model for Bulgarian/Macedonian, as we expect such approach to be beneficial since these two languages are rather similar. We plan to take a similar approach for Croatian and Serbian, hence why we do not train a single model for Croatian now, but wait until we have crawled Serbian data as well in the second release. We want to pay special attention to the training of Slovene systems, as we have native speakers of this language available in the consortium, allowing us to do a more in-depth study and evaluation. This takes some more time, meaning these results will come available in the following months. All code and models will be made publicly available.[1] Data sizes are shown in Table 2.1 for the MaCoCu data and other models we trained.

| Language | Corpora | Gigabytes | Tokens |
|---|---|---|---|
| Bulgarian | MaCoCu | 37 | 3,520,616,987 |
| Icelandic | MaCoCu | 4.4 | 687,996,096 |
| Macedonian | MaCoCu | 5.6 | 535,603,384 |
| Maltese | MaCoCu | 2.6 | 352,833,608 |
| | MaCoCu, Oscar, mC4 (clean) | 1.1 | 146,485,728 |
| | MaCoCu, Oscar, mC4 (all) | 3.2 | 439,481,778 |
| Turkish | MaCoCu | 35 | 4,388,022,947 |
| BERTovski (bg, mk) | MaCoCu, Oscar, mC4 | 74 | 7,026,841,189 |

**Table 2.1:** Data set sizes (bytes and tokens) for the MaCoCu monolingual release and all trained LMs.

Generally speaking, neural language models simply take a long time to train. Since only a month has passed since the first data release, a number of our models are still training. We plan to update the report periodically to present the new results of these models. The trained LMs are evaluated by *fine-tuning* them on downstream tasks. We use the same tasks across all five languages: Universal (UPOS) and language-specific (XPOS) part-of-speech tagging [8, 9, 10, 11], and Named Entity Recognition (NER). For Bulgarian, Macedonian and Turkish, we also evaluate on the Choice of Plausible Alternatives (COPA) task[2], which is a rather complex benchmark for causal reasoning. For UPOS/XPOS we use data from the Universal Dependencies project[3], while for NER we use the Wikiann data set [12]. There are two exceptions: for Macedonian POS-tagging we use the data sets of babushka-bench[4], while for Icelandic NER we use the MIM-GOLD-NER set [13]. The COPA data set was originally created for just English [14], so we translated it to Macedonian with help of a native speaker, while for Turkish the test set of COPA is available via the XCOPA data set [15]. We translated the Turkish training set and the Bulgarian training and test sets with Google Translate. In future work, we plan to perform full human translation of all the data sets with help of a professional translator.

### 2.1.1 Bulgarian/Macedonian

For Bulgarian and Macedonian, there was no high-quality monolingual model available yet. Likely, this makes it very useful that we train such a model from scratch. Therefore, we do not train on just the MaCoCu data. To create a model

[1] https://github.com/macocu/LanguageModels/

[2] https://people.ict.usc.edu/~gordon/copa.html

[3] https://universaldependencies.org/

[4] https://github.com/clarinsi/babushka-bench/

| RoBERTa | | Training | |
|---|---|---|---|
| Setting | Value | Setting | Value |
| LM-type | RoBERTa | Device | TPU |
| Architecture | RobertaForMaskedLM | Number of TPU cores | 8 |
| Masking | whole word | Max sequence length | 512 |
| Attention heads | 12 | Batch size | 32 |
| Hidden layers | 12 | Grad. accumulation steps | 8 |
| Hidden size | 768 | Total steps | 200,000 |
| Intermediate size | 3,072 | Warmup steps | 10,000 |
| Hidden activation | gelu | Peak learning rate | 5e-4 |
| Hidden dropout | 0.1 | Pad to max length | True |
| Layer norm eps | 1e-5 | Random seed | 2810 |
| Vocab size | 32,000 | | |

**Table 2.2:** Specific settings for training the BERTovski model from scratch. Other settings are left at their default value.

| | Bulgarian | | | | | | | Macedonian | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | UPOS | | XPOS | | NER | | COPA | UPOS | | XPOS | | NER | | COPA |
| Model | Dev | Test | Dev | Test | Dev | Test | Test | Dev | Test | Dev | Test | Dev | Test | Test |
| XLM-R (baseline) | 99.0 | 99.4 | 97.6 | 98.1 | 93.6 | 93.3 | 54.7 | 98.1 | 98.5 | 97.8 | 97.4 | 93.0 | 94.8 | 54.0 |
| BERTovski (50k) | 98.5 | 98.7 | 97.2 | 97.4 | 92.6 | 92.6 | — | 97.7 | 97.7 | 96.5 | 96.1 | 92.9 | 93.1 | — |
| BERTovski (100k) | 98.5 | 98.6 | 97.2 | 97.6 | 93.2 | 93.0 | — | 98.0 | 97.9 | 96.7 | 96.2 | 92.3 | 93.7 | — |
| BERTovski (150k) | 98.8 | 98.9 | 97.4 | 97.8 | 93.2 | 93.2 | — | 97.6 | 98.0 | 96.5 | 96.3 | 92.1 | 94.6 | — |
| BERTovski (200k) | 98.8 | 99.0 | 97.5 | 97.8 | 93.3 | 93.3 | 51.0 | 97.8 | 98.2 | 96.5 | 96.2 | 93.0 | 94.5 | 50.5 |
| XLM-R + BERTovski (20k) | 99.3 | 99.4 | 98.3 | 98.6 | 94.2 | 94.0 | 52.3 | 98.6 | 98.7 | 98.1 | 97.8 | 94.1 | 95.6 | 53.1 |

**Table 2.3:** Results for **Bulgarian** and **Macedonian** on the UPOS, XPOS, NER and COPA data sets.

that is most useful to the community, we also train on all Bulgarian and Macedonian data present in the Oscar [16] and mC4 [17] corpora. We train a RoBERTa-base model [5] from scratch, implemented in the Transformers library of HuggingFace [18]. We train our own vocabulary of 32,000 tokens[5], with a peak learning rate of 5e-4 and a batch size of 2,048.[6] Specific training settings are shown in Table 2.2. During training, we simply double the amount of Macedonian data to have a more balanced data division between the two languages. The final model, which we dubbed BERTovski, was trained for 200,000 steps or around 20 epochs, which took slightly over 2 weeks on a single v3 TPU.

Secondly, we train a model on the same data, but instead of training from scratch, we start from the large variant of multi-lingual XLM-Roberta (XLM-R, [19]). XLM-R is a state-of-the-art multi-lingual pretrained language model that is trained on 100 languages, including Bulgarian and Macedonian. The idea is that we start training from a model that has general knowledge about a lot of languages, which we then fine-tune on just our target language, in this case Bulgarian and Macedonian. This is likely a more efficient method of incorporating the language-specific data. We will apply this method also on Maltese, Turkish and Icelandic. For the continued XLM-R model, we keep the XLM-R-specific vocab of 250,000 pieces. We use a smaller learning rate (1e-4) and batch size (1,024), which is common practice when starting from an already trained model.

The results for Bulgarian and Macedonian are shown in Table 2.3. We observe that XLM-R (large) is a hard baseline to beat: it outperforms the models trained from scratch (BERTovski) for both Bulgarian and Macedonian. However, we do see that the scores for BERTovski (slightly) increase over time, which is promising. So far, the model is only trained on 200,000 steps, but we plan to train it for at least 500,000 steps. At that point, it might have caught up with XLM-R.

However, the good thing is that we already have a method to improve on XLM-R: if we continue training on its final checkpoint, we have improved performance on UPOS, XPOS and NER after only 20k steps, for both Bulgarian and Macedonian. On the COPA data set, we unfortunately do not see this improvement in performance. This is probably due to a deep semantic nature of the task (causal reasoning), which likely requires a much longer training of the language

[5] For training of the vocabulary we used an equal amount of Bulgarian and Macedonian data.

[6] Each of the 8 TPU cores has a batch size of 32 with 8 gradient accumulation steps, giving us an actual batch size of 2,048.

model. We will continue working on the problem for the second release.

Note that for continuing training from XLM-R, 20k steps is just a single epoch over the data. We look at these results as a proof-of-concept and expect more improvements once we train the models for longer. The XLM-R model is quite a bit larger than the RoBERTa-type models we train from scratch. This compared with the lower batch size resulted in a model that was about 4 times as slow to take a step, meaning 20k steps took 6 days on a v3 TPU. Now that we know this method works, we will experiment with different settings in order to best exploit the language-specific data. One promising method is to train a new vocabulary, instead of sticking with the large XLM-R vocab of 250,000 pieces.

### 2.1.2 Maltese

Similar as for Bulgarian and Macedonian, Maltese did not have high quality monolingual model already available. Therefore, we again train RoBERTa-type models from scratch on the Maltese data from the Oscar, mC4 and MaCoCu corpora. Since there is fewer data available for Maltese, the model does not have to be trained as long, allowing us to experiment a bit with the settings of the training process. In fact, we train three different systems:

- MaltBERT-clean: trained on data that is less likely to be machine-translated[7]
- MaltBERT-all: trained on all data that is classified as Maltese in the corpora, even if we suspect some of it to be machine translated
- MaltBERT-all-small: similar to MaltBERT-all, but with a smaller batch size and learning rate

We also trained a larger version of MaltBERT-all and MaltBERT-clean, but the training process failed after a few hours, with a loss that started increasing instead of decreasing. As there is fewer data available, we use a slightly smaller batch size (1,024) and learning rate (1e-4) than we did for the BERTovski models, but other settings are the same as in Table 2.2. The results are shown in Table 2.4. Interestingly, the model trained on the "clean" data is actually outperformed by the model on all data. Perhaps future work could look into training a model on even more data, by automatically translating an original English corpus into Maltese. It is also fascinating that XLM-R performs quite well on the Maltese tasks, while this language was not part of XLM-R's 100 languages. In other words, XLM-R never saw Maltese before the fine-tuning process, but can still do POS-tagging with around 95% accuracy. Once we continue training XLM-R on the Maltese data, we see clear improvements over just the XLM-R baseline, even after just 20k steps. Training for longer does not seem to be beneficial, as the model trained for 50k steps has similar performance.

| | **UPOS** | | **XPOS** | |
|---|---|---|---|---|
| **Model** | **Dev** | **Test** | **Dev** | **Test** |
| XLM-R (baseline) | 95.1 | 94.5 | 95.3 | 95.0 |
| MaltBERT-clean | 94.7 | 94.9 | 95.4 | 95.2 |
| MaltBERT-all | 95.4 | 95.5 | 95.6 | 95.9 |
| MaltBERT-all-small | 94.3 | 94.0 | 94.1 | 94.1 |
| XLM-R + MaltBERT (all, 20k) | 97.7 | 98.2 | 97.9 | 98.3 |
| XLM-R + MaltBERT (all, 50k) | 97.9 | 98.0 | 97.8 | 98.2 |

**Table 2.4:** Results for **Maltese** on the UPOS and XPOS data sets.

### 2.1.3 Icelandic

For Icelandic, we do not train a monolingual model from scratch, as there already exists a high quality monolingual LM for this language: IceBERT [20]. We consider it a waste of resources to train a similar model from scratch. We only continue training XLM-R again for 30k steps in a similar fashion as we did for the previous languages. The difference here is that we train on just the MaCoCu data, instead of on all Icelandic data available, so we can judge the impact of the MaCoCu data in a more controlled fashion. We can draw the conclusion that continuing training on just MaCoCu data is clearly beneficial: we improve performance over both the XLM-R and IceBERT baselines for all three tasks. In the future, we would like to examine performance when we continue training from IceBERT instead of XLM-R.

[7] We do this by excluding data from mt.{url}.com domains from the full data set, as in our experience, such domains often contain automatically translated data.

| | UPOS | | XPOS | | NER | |
|---|---|---|---|---|---|---|
| **Model** | **Dev** | **Test** | **Dev** | **Test** | **Dev** | **Test** |
| XLM-R (baseline) | 97.4 | 96.7 | 95.0 | 94.7 | 86.0 | 89.2 |
| IceBERT (baseline) | 96.4 | 95.9 | 93.9 | 93.6 | 85.3 | 89.6 |
| XLM-R + MaCoCu (30k) | 97.4 | 97.2 | 95.5 | 95.3 | 88.7 | 92.3 |

**Table 2.5:** Results for **Icelandic** on the UPOS, XPOS and NER data sets.

### 2.1.4 Turkish

For Turkish, we follow a similar process as for Icelandic, as there also exists a high quality monolingual LM for Turkish already: BERTurk [21]. We also continue training XLM-R for 30k steps on the monolingual Turkish MaCoCu data from the first release. The results are a bit less clear than before, since there is no clear improvement for the POS-tagging tasks. However, for NER we clearly improve over both the XLM-R and BERTurk baselines. We also show performance for models fine-tuned on the machine-translated COPA training data into Turkish. However, COPA does have a human translated test set [22], allowing us to compare performance on two test sets – the machine-translated (MT) and the human-translated (HT) one. First, we actually do see improved performance when we continue training from XLM-R, for both the HT and the MT versions. Second, we see improved performance for all models on the human translated test set. Even if the model was still only trained on translations, it helps to see higher quality text during evaluation. This is promising, since we plan to have the COPA sets for the other languages translated by professional translators.

| | UPOS | | XPOS | | NER | | COPA | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Dev** | **Test** | **Dev** | **Test** | **Dev** | **Test** | **Test (MT)** | **Test (HT)** |
| XLM-R (baseline) | 89.1 | 89.4 | 90.7 | 90.6 | 93.3 | 93.5 | 54.4 | 54.8 |
| BERTurk (baseline) | 88.1 | 88.3 | 89.7 | 89.4 | 93.3 | 93.2 | 53.6 | 54.2 |
| XLM-R + MaCoCu (30k) | 89.5 | 89.4 | 90.7 | 90.5 | 94.2 | 94.2 | 55.4 | 56.0 |

**Table 2.6:** Results for **Turkish** on the UPOS, XPOS, NER and COPA data sets. The COPA test sets are either machine translated (MT) or human translated (HT).

# 3 Evaluation of bilingual corpora

We evaluate the quality of the bilingual corpora in two ways: (i) automatically, by training neural machine translation (NMT) systems with and without MaCoCu's parallel data and (ii) manually, by having professional translators judge if the translations of the NMT system that includes MaCoCu's data are of higher quality data than those produced by the baseline NMT system, which does not include MaCoCu's data. These two evaluation methods are described below.

## 3.1 Automatic evaluation of bilingual corpora

We build MT systems from English into six languages targeted in MaCoCu's first release. Our main aim is to show that we can achieve higher performance by also training an MT system on the data of the first release of MaCoCu. For each of the languages, we train a strong baseline system that is trained on large parallel data sets that were already available.[1] We then compare its performance to a system with the exact same architecture that is trained on the same data plus the parallel data of MaCoCu's first release. We hypothesise that the latter achieves better performance.

The MT system we use is a Transformer [23] model implemented in Marian [24], with settings previously used in the ParaCrawl [25] evaluation of parallel corpora. We train a Transformer-base model with 6 layers for the encoder and decoder and 8 attention heads, with a hidden size of 2,048. For each language, we train a vocab of 32,000 pieces through byte-pair encoding [26, 27]. We truncate the input to a maximum of 200 of such pieces. During training, we automatically use a batch size that fits into our memory (32GB on a GPU). We use a learning rate of 0.0003, with a warm-up of 16,000 steps. During training, we apply label smoothing with a value of 0.1. Training is stopped using early stopping, calculated with BLEU after each epoch, with a patience of three. We use the same settings in all our experiments.

To evaluate performance, we use the following well-established MT metrics: COMET [2], BLEU [3], CHRF [28], BERT-score [29] and BLEURT [30]. For brevity, we only show performance of the traditionally most important metric in MT (BLEU), as well as the current best performing metric (COMET).

We train the baseline systems for each language on the CommonCrawl [31], ParaCrawl [25] and Tilde [32] data sets. Since there can be overlap between the corpora, we deduplicate the final training sets based on the source sentences. Data set statistics are shown in Table 3.1. Note that due to the deduplication, the columns *Baseline size* and *MaCoCu size* do not sum to *Total size*.

| Language | Baseline corpora | Baseline size | MaCoCu size | Total size |
|---|---|---:|---:|---:|
| Bulgarian | ParaCrawl, CommonCrawl | 14,424,709 | 2,326,861 | 16,245,592 |
| Croatian | ParaCrawl, CommonCrawl, Tilde | 5,106,096 | 2,114,409 | 7,029,981 |
| Icelandic | ParaCrawl, CommonCrawl, Tilde | 3,281,592 | 349,246 | 3,537,840 |
| Maltese | ParaCrawl, Tilde | 2,291,030 | 1,056,735 | 3,175,333 |
| Slovene | ParaCrawl, CommonCrawl, Tilde | 10,870,935 | 2,337,395 | 12,741,147 |
| Turkish | ParaCrawl | 4,927,668 | 4,709,457 | 9,575,968 |

**Table 3.1:** Data set sizes (sentence pairs) for the baseline corpora and the MaCoCu corpora per language.

We evaluate performance on the data sets Flores (devtest, [1]), TED [33], WikiMatrix [34], WMT [35, 36, 37] and QED [38]. Again, we only show the performance on Flores for brevity, but scores on all data sets are available upon request. Generally, relative performance across data sets was very similar. Our main results are shown in Table 3.2. We clearly improve performance by using the new MaCoCu data on all languages, except for Turkish. Even for Icelandic, for which we crawled relatively little data, we still find a considerable improvement in terms of the COMET score.

The score for Turkish is somewhat unexpected, since this is the language we actually crawled the most data for. The crawled MaCoCu data for this language is either of lower quality, or does not fit the Flores domain. To determine which is more likely, we evaluate on all evaluation sets for Turkish and show the results in Table 3.3. This paints a very different picture, with the MaCoCu model clearly improving over the baseline in almost all evaluation sets. This tells us that

[1] Data selected from: https://opus.nlpl.eu/

perhaps the MaCoCu data is in quite a different domain than the Flores data (which is based on Wikipedia paragraphs), but that the data is not of lower quality. Nevertheless, we will pay special attention to the Turkish crawling and cleaning process in the next release.

| | **Bulgarian** | | **Croatian** | | **Icelandic** | | **Maltese** | | **Slovene** | | **Turkish** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CO** | **BL** | **CO** | **BL** | **CO** | **BL** | **BS** | **BL** | **CO** | **BL** | **CO** | **BL** |
| Baseline | 79.2 | 41.0 | 80.6 | 29.6 | 47.0 | 23.5 | 80.7 | 37.6 | 77.0 | 29.8 | 81.3 | 28.2 |
| Baseline + MaCoCu | **+0.9** | **+0.3** | **+0.5** | **+0.2** | **+2.4** | **+0.4** | **+1.0** | **+2.1** | **+0.3** | **+0.1** | **-1.2** | **+0.2** |

**Table 3.2:** Performance on Flores devtest for our two systems. CO, BL and BS are short for COMET, BLEU and BERT-Score.

| | **Flores** | | **WMT16** | | **WMT17** | | **WMT18** | | **TED** | | **Wiki** | | **QED** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **CO** | **BL** | **CO** | **BL** | **CO** | **BL** | **BL** | **BL** | **CO** | **BL** | **CO** | **BL** | **CO** | **BL** |
| Baseline | 81.3 | 28.2 | 72.6 | 20.3 | 77.3 | 21.4 | 74.0 | 18.9 | 62.9 | 17.1 | 61.1 | 22.8 | 51.1 | 14.0 |
| Baseline + MaCoCu | **-1.2** | **+0.2** | **+2.3** | **+0.7** | **+0.4** | **+0.7** | **+2.2** | **+0.6** | **+1.6** | **-0.1** | **+1.0** | **+0.6** | **+0.8** | **-0.2** |

**Table 3.3:** Performance for our two **Turkish** systems on all evaluation sets. CO and BL are short for COMET and BLEU.

Generally speaking, the results are very promising: training on additional MaCoCu data improves performance across languages, evaluation sets and evaluation metrics. We conclude that the crawled data is of high quality and adds value to the MT research community. However, we are also interested in whether this crawled data is of *higher* quality than the data crawled in ParaCrawl. To test this, we perform a controlled experiment in which we limit the ParaCrawl data to exactly the same amount of sentences as the MaCoCu data.[2] Icelandic and Turkish are not part of this evaluation, as the former did not have enough data available to train a high quality model, while the latter was not part of ParaCrawl. The results are shown in Table 3.4 and are a bit mixed. For Bulgarian and Croatian, the ParaCrawl has better performance, while for Slovene it is actually the MaCoCu model. In any case, from this experiment we cannot conclude that the MaCoCu data is of higher quality than ParaCrawl.

| | **Bulgarian** | | **Croatian** | | **Maltese** | | **Slovene** | |
|---|---|---|---|---|---|---|---|---|
| | **CO** | **BL** | **CO** | **BL** | **BS** | **BL** | **CO** | **BL** |
| ParaCrawl | 68.4 | 35.5 | 72.9 | 26.8 | 80.4 | 35.4 | 67.2 | 26.4 |
| MaCoCu | **-2.9** | **-1.4** | **-1.4** | **-0.5** | **+0.3** | **0.4** | **+2.9** | **+0.2** |

**Table 3.4:** Performance on Flores devtest for systems either trained on just MaCoCu data or just on ParaCrawl data. The ParaCrawl data is limited to be the exact same size as the MaCoCu data (in sentences). CO, BL and BS are short for COMET, BLEU and BERT-Score.

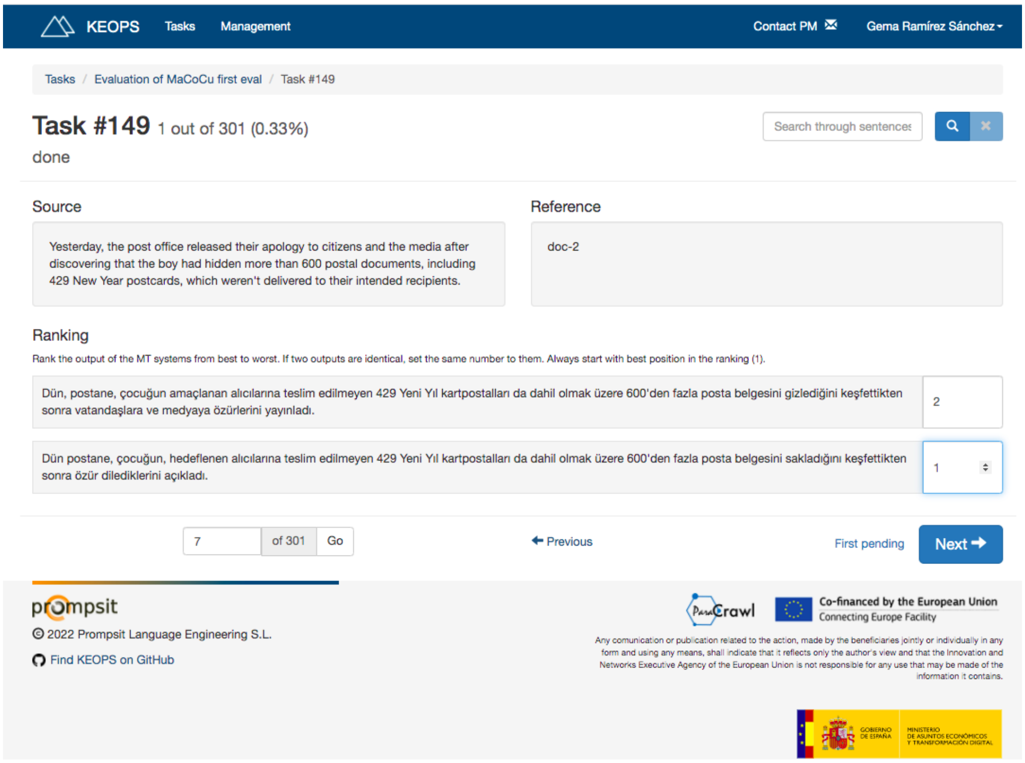
## 3.2 Human evaluation of bilingual corpora

We perform a human evaluation of the performance of our two systems as described in Section 3.1: the best (baseline) model without MaCoCu and the best model with MaCoCu data added. The aim is to check whether the increased performance of the MaCoCu models (as shown in Table 3.2) is confirmed by professional translators.

Since we compare two systems per language, we set up a relative ranking task for each language. We randomly selected a subset of the Flores data set of 301 sentences, which all had the same English source sentence for each of our target languages. Annotators are given the source sentence plus the two outputs of our two MT systems (not knowing which is which) and have to determine whether they prefer the first translation, the second translation, or if there is no difference in quality. We hired 3 annotators per language, for a total of 18 annotators, through the *Translated* company. They were all selected among professional translators with experience in MT evaluation.

[2] Limiting ParaCrawl to the same amount of bytes showed very similar results.

Annotations were performed using the KEOPS online tool,[3] which implements, among other tasks, relative ranking. In the annotation screen we show only the source sentence, information about document boundaries and the two MT outputs that annotators are asked to rank. Besides the raw annotations, KEOPS computes time spent on each instance and on the whole task on average and outputs a chart with final results once the task is completed. We provide some screenshots of the annotation screen and task recap upon completion in Figures 3.1 and 3.2 below.



Task #149 1 out of 301 (0.33%)

done

Source

Yesterday, the post office released their apology to citizens and the media after discovering that the boy had hidden more than 600 postal documents, including 429 New Year postcards, which weren't delivered to their intended recipients.

Reference

doc-2

Ranking

Rank the output of the MT systems from best to worst. If two outputs are identical, set the same number to them. Always start with best position in the ranking (1).

Dün, postane, çocuğun amaçlanan alıcılarına teslim edilmeyen 429 Yeni Yıl kartpostalları da dahil olmak üzere 600'den fazla posta belgesini gizlediğini keşfettikten sonra vatandaşlara ve medyaya özürlerini yayınladı. 2

Dün postane, çocuğun, hedeflenen alıcılarına teslim edilmeyen 429 Yeni Yıl kartpostalları da dahil olmak üzere 600'den fazla posta belgesini sakladığını keşfettikten sonra özür dilediklerini açıkladı. 1

7 of 301 Go

© 2022 Prompsit Language Engineering S.L.

Co-financed by the European Union
Connecting Europe Facility

Any comunication or publication related to the action, made by the beneficiaries jointly or individually in any form and using any means, shall indicate that it reflects only the author's view and that the Innovation and Networks Executive Agency of the European Union is not responsible for any use that may be made of the information it contains.

**Figure 3.1:** Annotation screen in KEOPS for the relative ranking of English to Turkish MT outputs from human evaluators.

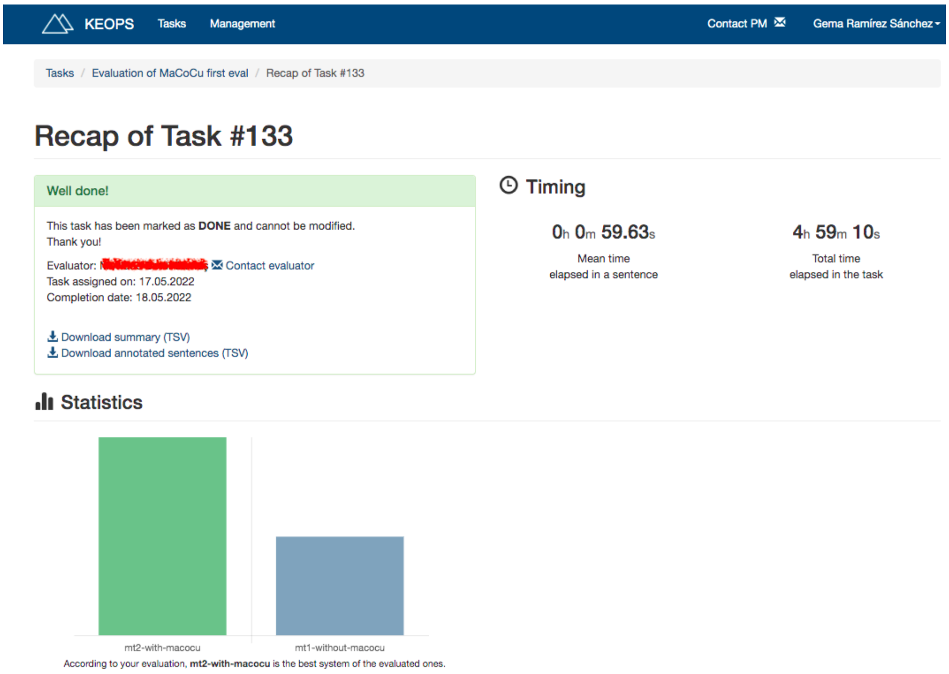
The main result of the annotation process is shown in the top half of Table 3.5. We find that, for each language, annotators more often prefer the model with MaCoCu data (third row with results) over the model without MaCoCu data (top row with results). This even includes Turkish, for which the automatic metrics showed more of a mixed result (Table 3.3).

| | bg | hr | is | mt | sl | tr |
|---|---|---|---|---|---|---|
| Without MaCoCu preferred (%) | 9.2 | 25.7 | 22.9 | 26.1 | 37.4 | 36.2 |
| Same quality (%) | 79.8 | 43.9 | 49.5 | 36.2 | 22.9 | 22.3 |
| With MaCoCu preferred (%) | 11.0 | 30.5 | 27.6 | 37.7 | 39.6 | 41.5 |
| Exact same translation (%) | 55.5 | 14.0 | 5.6 | 9.0 | 17.3 | 6.6 |
| Average time per instance (seconds) | 37.6 | 44.6 | 53.2 | 38.1 | 33.2 | 83.1 |
| Hard disagreements (%) | 1.8 | 8.7 | 7.4 | 13.2 | 16.5 | 19.0 |
| Inter-annotator agreement | 0.30 | 0.35 | 0.18 | 0.34 | 0.60 | 0.39 |
| P-value of significance test | 0.641 | 0.162 | 0.183 | **0.001** | 0.549 | 0.126 |

**Table 3.5:** Preferences in percentages for the annotations of the subset of 301 sentences of the Flores devtest set. For each language, the values are averaged over the three annotators. Inter-annotator agreement is calculated by Cohen's Kappa.

In the bottom half of Table 3.5 we show some more detailed results. For each language, we show the percentage of times

[3] https://keops.prompsit.com/



Tasks / Evaluation of MaCoCu first eval / Recap of Task #133

Recap of Task #133

Well done!

This task has been marked as **DONE** and cannot be modified.
Thank you!

Evaluator: [redacted] Contact evaluator
Task assigned on: 17.05.2022
Completion date: 18.05.2022

Download summary (TSV)
Download annotated sentences (TSV)

Timing

0h 0m 59.63s
Mean time elapsed in a sentence

4h 59m 10s
Total time elapsed in the task

Statistics

According to your evaluation, **mt2-with-macocu** is the best system of the evaluated ones.

**Figure 3.2:** Recap of the task for one of the English to Turkish MT outputs relative ranking annotations, with clear preference for the MT system including MaCoCu data.

the two MT systems output the exact same sentence. We see that this was especially often the case for Bulgarian (55.5%), which explains the high percentage of annotators indicating that the translations were of the same quality. The sixth row of results shows the percentage of "hard disagreements" which we define as the number of time annotator *A* prefers the output of system *X*, while annotator *B* prefers the output of system *Y*. In other words, if one of the annotators ranks the translations the same, it is by definition not a hard disagreement. The percentages are averaged over the three pairs of annotators per language. We observe that annotators do not have hard disagreements often: less than 20% of the time for all languages.

The seventh row of Table 3.5 shows the inter-annotator agreement, as calculated by Cohen's kappa. Annotators are generally in agreement, though the score for Icelandic is on the lower side. In the eighth row of results, we show the p-value of applying a binomial t-test on the concatenated annotations of the three translators per language. Even though annotators preferred the MT system with MaCoCu data for all six languages, we only find a significant difference for Maltese. For Croatian, Icelandic and Turkish we are perhaps approaching significance, while for Bulgarian and Slovene we are further off. This is not surprising, as the latter two languages had by far the most parallel data already available (14.4M and 10.9M sentences, see Table 3.1), which makes it harder to significantly improve its performance.

## 3.3 Conclusion

In this section we evaluated the parallel corpora in the first MaCoCu release. We showed that training on additional MaCoCu data resulted in more accurate neural machine translation systems, as compared to non-MaCoCu baselines, when looking at automatic evaluation metrics. This result held up across different languages, evaluation sets and evaluation metrics. We then confirmed this result in a human evaluation study, in which professional translators also generally preferred the translations of the MaCoCu-trained models over the non-MaCoCu baselines.

# Bibliography

[1] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 05 2022. [Online]. Available: https://doi.org/10.1162/tacl_a_00474

[2] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for mt evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2685–2702.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[6] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, "Bertje: A dutch bert model," *arXiv preprint arXiv:1912.09582*, 2019.

[7] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a tasty French language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7203–7219. [Online]. Available: https://aclanthology.org/2020.acl-main.645

[8] K. Simov, P. Osenova, A. Simov, and M. Kouylekov, "Design and implementation of the bulgarian hpsg-based treebank," *Research on Language and Computation*, vol. 2, no. 4, pp. 495–522, 2004.

[9] S. Čéplö, "Constituent order in maltese: A quantitative analysis," 2018.

[10] t. Arnardóttir, H. Hafsteinsson, E. F. Sigurdhsson, K. Bjarnadóttir, A. K. Ingason, H. Jónsdóttir, and S. Steingrímsson, "A Universal Dependencies conversion pipeline for a Penn-format constituency treebank," in *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*. Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 16–25. [Online]. Available: https://www.aclweb.org/anthology/2020.udw-1.3

[11] U. Türk, F. Atmaca, Ş. B. Özateş, G. Berk, S. T. Bedir, A. Köksal, B. Ö. Başaran, T. Güngör, and A. Özgür, "Resources for turkish dependency parsing: Introducing the boun treebank and the boat annotation tool," *arXiv preprint arXiv:2002.10416*, 2020.

[12] A. Rahimi, Y. Li, and T. Cohn, "Massively multilingual transfer for NER," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 151–164. [Online]. Available: https://aclanthology.org/P19-1015

[13] S. L. Ingólfsdóttir, Á. A. Gudhjónsson, and H. Loftsson, "MIM-GOLD-NER – named entity recognition corpus (21.09)," 2020, CLARIN-IS. [Online]. Available: http://hdl.handle.net/20.500.12537/140

[14] M. Roemmele, C. A. Bejan, and A. S. Gordon, "Choice of plausible alternatives: An evaluation of commonsense causal reasoning." in *AAAI spring symposium: logical formalizations of commonsense reasoning*, 2011, pp. 90–95.

[15] E. M. Ponti, G. Glavaš, O. Majewska, Q. Liu, I. Vulić, and A. Korhonen, "XCOPA: A multilingual dataset for causal commonsense reasoning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2362–2376. [Online]. Available: https://aclanthology.org/2020.emnlp-main.185

[16] P. J. Ortiz Su'arez, L. Romary, and B. Sagot, "A monolingual approach to contextualized word embeddings for mid-resource languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1703–1714. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.156

[17] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

[20] V. Snæbjarnarson, H. B. Símonarson, P. O. Ragnarsson, S. Ingólfsdóttir, H. P. Jónsson, V. Thorsteinsson, and H. Einarsson, "A warm start and a clean crawled corpus–a recipe for good language models," *arXiv preprint arXiv:2201.05601*, 2022.

[21] S. Schweter, "Berturk - bert models for turkish," Apr. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3770924

[22] E. M. Ponti, G. Glavaš, O. Majewska, Q. Liu, I. Vulić, and A. Korhonen, "XCOPA: A multilingual dataset for causal commonsense reasoning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2362–2376. [Online]. Available: https://aclanthology.org/2020.emnlp-main.185

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, 2018. [Online]. Available: https://arxiv.org/abs/1804.00344

[25] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza, "ParaCrawl: Web-scale acquisition of parallel corpora," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4555–4567. [Online]. Available: https://aclanthology.org/2020.acl-main.417

[26] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[27] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[28] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 392–395. [Online]. Available: https://aclanthology.org/W15-3049

[29] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2019.

[30] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7881–7892. [Online]. Available: https://aclanthology.org/2020.acl-main.704

[31] A. El-Kishky, V. Chaudhary, F. Guzmán, and P. Koehn, "CCAligned: A massive collection of cross-lingual web-document pairs," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language

*Processing (EMNLP 2020)*. Online: Association for Computational Linguistics, November 2020, pp. 5960–5969. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.480

[32] R. Rozis and R. Skadiņš, "Tilde MODEL - multilingual open data for EU languages," in *Proceedings of the 21st Nordic Conference on Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, May 2017, pp. 263–265. [Online]. Available: https://aclanthology.org/W17-0235

[33] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. [Online]. Available: https://arxiv.org/abs/2004.09813

[34] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, and F. Guzmán, "WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 1351–1361. [Online]. Available: https://aclanthology.org/2021.eacl-main.115

[35] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Névéol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri, "Findings of the 2016 conference on machine translation," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 131–198. [Online]. Available: https://aclanthology.org/W16-2301

[36] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. Specia, and M. Turchi, "Findings of the 2017 conference on machine translation (WMT17)," in *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 169–214. [Online]. Available: https://aclanthology.org/W17-4717

[37] O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, P. Koehn, and C. Monz, "Findings of the 2018 conference on machine translation (WMT18)," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 272–303. [Online]. Available: https://aclanthology.org/W18-6401

[38] A. Abdelali, F. Guzman, H. Sajjad, and S. Vogel, "The AMARA corpus: Building parallel language resources for the educational domain," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 1856–1862. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/877_Paper.pdf