

# Benchmarking Azerbaijani Neural Machine Translation

Chih-Chen Chen<sup>1</sup>, William Chen<sup>1</sup>

<sup>1</sup>University of Central Florida

## Abstract

Little research has been done on Neural Machine Translation (NMT) for Azerbaijani. In this paper, we benchmark the performance of Azerbaijani-English NMT systems on a range of techniques and datasets. We evaluate which segmentation techniques work best on Azerbaijani translation and benchmark the performance of Azerbaijani NMT models across several domains of text. Our results show that while Unigram segmentation improves NMT performance and Azerbaijani translation models scale better with dataset quality than quantity, cross-domain generalization remains a challenge.

## 1. Introduction

With the recent growth in online resources, robust NLP systems have become increasingly available for many of the world's languages. However, this growth has not been enjoyed equally and technologies for many languages are still under-developed, especially relative to the size of their speaker population. This remains the case for morphologically-complex languages, which have been considered a challenge for NLP systems due to the frequency of rare/unknown words. One such example is Azerbaijani, a Turkic language with a highly agglutinative and complex morphology. It has two major varieties: the Northern variant is spoken in the Republic of Azerbaijan, while Southern Azerbaijani regions of Iran. Our experiments focus on Northern Azerbaijani, which is written in Latin script and has considerably more online resources that are able to support the development of NMT systems.

Little work has been done on NLP systems for Azerbaijani, and even less on machine translation and other generative Seq2Seq tasks. Specifically, there is a lack of benchmarks on the performance of Azerbaijani NMT and the methods that could be used to improve it. Existing studies either include private datasets with unpublished training, testing, and validation splits [1] or solely evaluate on very low-resource scenarios with transfer learning techniques [2]. We build off the approach developed by Guntara et al. [3], who sought to develop benchmarks for Indonesian NMT, and extend it to include the evaluation of different pre-processing techniques for Azerbaijani NMT. Our goal is to help address these problems by investigating the following research questions regarding Azerbaijani translation:

1. What segmentation methods work best for Azerbaijani NMT?


---

*The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), June 7-8, Koper, Slovenia*

✉ chihchen.chen@outlook.com (C. Chen); wchen6255@knights.ucf.edu (W. Chen)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. How important is data cleanliness versus training corpora size for Azerbaijani NMT?
3. How do Azerbaijani translation systems perform across different language domains?

To answer these questions, we set up the following experiments:

1. We evaluate the performance of different segmentation algorithms to see which perform best for Azerbaijani.
2. We evaluate the effectiveness of scaling to larger training corpora at the cost of alignment quality in Azerbaijani NMT.
3. We categorize open-source Azerbaijani corpora into different domains and evaluate the effectiveness of NMT models trained on individual and multiple domains.

Our results showed that both the choice of evaluation metric and segmentation algorithm have a large impact in determining which models are the best performing, showing the importance of evaluating across multiple metrics. We also found that sentence alignment quality was a large factor in model performance; the addition of large but noisy/out-of-domain training datasets did not necessarily translate to improved performance.

## 2. Related Work

Studies on morphologically-complex languages tend to focus on the higher-resource Turkish or extremely low-resource languages like Inuktitut or Quechua. However, there have been many experiments that use Azerbaijani to demonstrate the effects of transfer learning and multilinguality due to its relationship with Turkish. Early MT systems for Azerbaijani were built by Fatullayev et al. [4]. Their models were based off of a hybrid between rule-based and statistical machine translation, and could translate to/from English and Turkish. Qi et al. [2] experimented with Azerbaijani in a low-resource setting to improve NMT with aligning pre-trained word embeddings. They showed that including Turkish with Azerbaijani in multilingual NMT significantly improved BLEU score. Neubig and Hu [5] explored training paradigms for multilingual NMT that also leverage Turkish to improve Azerbaijani translation. Kim et al. [6] showed the effectiveness of cross-lingual word-embeddings in improving low-resource Azerbaijani NMT. The most recent work on bilingual Azerbaijani NMT was by Maimaiti et al. [1], who used Azerbaijani and Uzbek to Chinese translation as case studies for transfer learning with pre-trained lexicon embeddings.

Many studies have been done on the effect on subword segmentation algorithms on downstream NMT. Sennrich et al. [7] and Kudo [8] show that such algorithms improve the performance of NMT models using Byte-Pair Encoding (BPE) and Unigram segmentation respectively. While BPE has generally been the standard, recent works show that the Unigram algorithm performs better on agglutinative languages [9][10][11]. Mager et al. [12] compared the performance of BPE to morphological segmentation algorithms for indigenous American languages and found that SOTA morphological segmentation methods did not translate to improved performance on NMT. Results in a similar study by Sälevä and Lignos [13] were inconclusive when comparing BPE with LMVR [14] and MORSEL [15] on Nepali, Sinhala, and Kazakh; the best performing segmentation algorithm was language dependent and the results were statistically

indistinguishable. Pre-processing techniques have also been a feature of interest in low-resource translation shared tasks. Chen and Fazio [16] found that Unigram segmentation [8] performed the best for Marathi-English translation at LoResMT 2021 [17]. Vázquez et al. [18] leveraged data cleaning and normalization techniques to overcome differences in orthographic conventions for multilingual models at AmericasNLP 2021 [19].

### 3. Experimental Setup

For all of our experiments we use the OpenNMT-py [20] implementation of the Transformer [21]. We use the set-up from Chen and Fazio [9], which has been shown to perform well with agglutinative languages. The architecture is comprised of 6 encoder/decoder layers, 8 attention heads, size 256 word vectors, and a feed-forward dimension of 2048. The models were trained for 50,000 steps with a batch size of 32.

Translation quality is evaluated using COMET [22] and the sacreBLEU [23] implementations of BLEU [24] and chrF [25] scores. Kocmi et al. [26] recommended the use of COMET and chrF, which they found were the metrics that best correspond to human judgement. We also provide BLEU scores due to its standard use in machine translation. Each model was independently trained 10 times such that the presented scores below are the average across all trials.

#### 3.1. Q1: Segmentation Algorithms for Azerbaijani

A common pre-processing technique to improve the performance of NLP systems is subword segmentation: separating words into small units to decrease vocabulary size and help the model generalize to unknown vocabulary. The goal of our first set of experiments is to identify which subword segmentation algorithms work best for Azerbaijani. We use the Azerbaijani-English portion of WikiMatrix [27], which consists of 276k parallel sentences. The WikiMatrix dataset provides the LASER [28] score of each sentence pair, which measures the likelihood of a sentence pair being mutual translations. Filtering out sentences with a score less than 1.04 (the recommended LASER threshold) reduces the dataset size to 70,725. The cleaned dataset is then split into 47,385 training sentences, 11,670 validation sentences, and 11,670 test sentences.

Models are trained on text segmented by different techniques: Byte-Pair Encoding (BPE) [7], BPE-Guided [29], Unigram [8], and PRPE [30]. BPE and Unigram segmentation are the two most popular segmentation algorithms used in state-of-the-art NMT systems due to their flexibility and ease of use. BPE-Guided [29] and PRPE [30] are morphologically-motivated algorithms that were shown to perform well on NMT for agglutinative languages [29][9]. Prior to subword segmentation, the text is first tokenized by Moses Tokenizer [31].

BPE first splits the corpus into a character level representation. The most frequently occurring pair of tokens are then merged together, a process that is repeated until a pre-defined number of merge operations have been reached. BPE-Guided is an extension of the BPE algorithm that incorporates morphological information through a list of known affixes. BPE-Guided creates a glossary of words that do not contain any known affixes, which is then used by the main BPE algorithm as a list of words to not segment.

Unigram segmentation is a probabilistic segmentation algorithm based on a unigram language model [8]. A vocabulary of a pre-defined size is first built by only keeping subwords that least

reduce the loss of calculating subword occurrence probabilities via the expectation-maximization algorithm. The output segmentation of a word is then obtained by choosing the most probable segmentation candidate obtained from the Viterbi algorithm [32].

Prefix-Root-Postfix-Encoding (PRPE) segments a word into three main parts: a prefix, root and a postfix. The algorithm first learns a subword vocabulary of prefixes and postfixes with the help of a language-specific heuristic. PRPE then uses any detected instances of those affixes in a word to extract potential roots and obtain the most probable segmentation of the word.

Segmentation Algorithm	BLEU	chrF	COMET	<i>p</i> -value
None	1.596	13.136	-1.205	
BPE	1.567	13.710	-1.207	0.0240
BPE-Guided	1.517	12.010	-1.234	0.0006
PRPE	1.625	13.615	-1.195	0.0099
Unigram	1.730	14.150	-1.188	0.0013

**Table 1**

A comparison of different segmentation algorithms on Northern Azerbaijani to English NMT. Higher scores indicate better performance. *p*-values are calculated using the average COMET score of the given algorithm compared to that of no segmentation.

The BLEU, chrF, and COMET scores are found in Table 1; *p*-values calculated with a paired Student’s t-test between a chosen segmentation algorithm’s COMET score and the no segmentation baseline are also included. Almost all segmentation methods obtained higher chrF and BLEU scores than the no segmentation baseline. Unigram segmentation performed the best, achieving the highest scores in all three evaluation metrics. PRPE was the second best performing algorithm in BLEU and COMET, but scored lower than BPE in terms of chrF. Interestingly, these two algorithms were also the only ones that performed better than the baseline in terms of COMET score. These results show that both the metric and segmentation algorithm used can have a significant impact on what models are designated as "the best performing", and further encourage the reporting of across multiple evaluation metrics in future work.

### 3.2. Q2: Dataset Size vs Cleanliness

We conducted a second set of experiments to examine the tradeoff between dataset cleanliness and dataset size in regards to NMT performance by using the alignment scores provided by the WikiMatrix dataset [28] as a measurement of cleanliness. To do so, we created additional training datasets with the WikiMatrix sentence pairs left unused in Section 3.1. We combine these remaining sentences with the clean 47k sentence training set to form a noisy 252k sentence training dataset. As a middle ground, we also create a third training dataset of 120k sentences by only keeping sentence pairs with a score of at least 1.03 from the large noisy dataset. The validation and test sets are reused from 3.1. The text was not pre-processed with any subword segmentation algorithm to isolate any impact on the performance metrics to the change in training data.

The results (Table 2) provide an interesting reflection of how the evaluation metrics are

Training Dataset	# Sentences	BLEU	chrF	COMET
Clean (T=1.04)	47,385	1.596	13.136	-1.205
Slightly Noisy (T=1.03)	119,725	2.276	12.614	-1.292
Noisy (T=0)	252,255	2.488	11.460	-1.399

**Table 2**

A comparison of the tradeoff between dataset size and cleanliness. T is the LASER score threshold use to filter sentence pairs, which is a measurement of the likelihood that two sentences are mutual translations.

calculated. BLEU [24] scores increased as the training dataset size grew, but chrF [25] and COMET [22] scores decreased. We hypothesize that this is because the additional training data increased the vocabulary size of the model and thus allowed it to recognize otherwise unknown words in the test set. Our results corroborate the findings of Kocmi et al. [26] and show the inaccuracy of BLEU compared to other metrics: evaluating only with BLEU would indicate that training on the smaller dataset was worse despite the opposite holding true.

### 3.3. Q3: Domain Benchmarks

Our final experiment was to evaluate the performance of an Azerbaijani NMT model across several domains of text. We first obtained all Azerbaijani-English (az-en) data from OPUS [33], which consist of the following parallel corpora: WikiMatrix [27], CCMatrix [34], Tatoeba, ELRC public corpora, Tanzil, GNOME [35], QED [36], TED2020 [37], and XLEnt [38]. The corpora were categorized by domain, of which the domains with little data (lecture, news, and tech) were aggregated into a larger “Mixed” domain dataset. We thus evaluate the model on four different datasets: General (1,325,660 lines), Religious (269,445 lines), Entities (298,236 lines), and Mixed (68,256 lines). Each dataset was then split into 66.7% training sentences, 16.6% validation sentences, and 16.6% test sentences. All text is pre-processed with Moses Tokenizer [31] and segmented with a Unigram segmentation model [8].

Dataset	# Sentences	Domain
CCMatrix	1,251,255	General
WikiMatrix (T=1.04)	70,725	General
Tatoeba	3,680	General
ELRC	129	News
Tanzil	269,445	Religious
GNOME	40,075	Tech
QED	16,442	Lecture
TED2020	11,610	Lecture
XLEnt	298,236	Entities
<b>Total</b>	<b>1,961,597</b>	

**Table 3**

Dataset Statistics

We independently train models on each dataset. To evaluate the system’s ability to generalize across domains, we train another model on the data combined across all 4 datasets. The models are trained for 300,000 steps and are evaluated using the best performing checkpoint on the validation set. The 4 domain-specific models are evaluated on the test set of their domain and the model trained on combined data is evaluated on each domain.

Test Set	Trained on Domain Only			Trained on Combined Data		
	BLEU	chrF	COMET	BLEU	chrF	COMET
General	5.55	16.999	-1.069	3.981	14.795	-1.1658
Religious	23.199	44.535	-0.818	17.285	34.285	-0.6010
Entities	7.607	19.845	-0.929	1.279	11.428	-1.1751
Mixed	22.725	35.648	-0.136	4.555	15.293	-1.0216

**Table 4**

A comparison of the BLEU, chrF, and COMET scores between models trained on a specific data domain and a model trained on data across all domains.

Most of the domain-specific models performed better than the model trained on combined data (Table 4). An exception was on the Religious dataset; while the Religious model performed better than the Combined Data model in terms of BLEU and chrF, the Combined Data model achieved a better COMET score. This indicates that training on a more general dataset allowed the model to output more words that were closer to the label translation in the embedding space (higher COMET score) but differed in terms of subwords/characters used (lower BLEU and chrF score). These results also corroborate those of 3.2, again showing the importance of data cleanliness. Models trained on the smaller and cleaner Religious and Mixed datasets performed better than those trained on the larger General, Entities, and Combined datasets. The result is particularly noticeable with the Mixed dataset model, which achieved a COMET score of -0.136 despite having only 45,500 training sentences.

## 4. Conclusion

We trained several Azerbaijani NMT models on text segmented by different algorithms and show that using Unigram segmentation can noticeably improve translation quality. We also demonstrate that properly cleaning data can lead to significant gains in performance, even when shrinking the training corpora. Finally, we evaluated the performance of Azerbaijani-English NMT models across multiple domains. Our results demonstrate that while generalizing across domains remains a challenge for Azerbaijani NMT, specialized models are still able to achieve a competitive performance.

## 5. Future Work

Our experiments focused only on Northern Azerbaijani due to scarcity of data for the Southern variant. One route for exploration to develop NMT systems for the latter is to compare the effectiveness of lower-resource cross-dialectal transfer from Northern Azerbaijani against

higher-resource cross-lingual transfer from Turkish. Developing NMT systems for Southern Azerbaijani is particularly challenging since it is written in Arabic script, introducing the need for transliteration to properly take advantage of transfer learning paradigms. Further evaluation could also be done on the transfer learning and multilingual techniques used to improve Azerbaijani translation introduced in previous works. While those studies show that such techniques are able to improve translation quality over a simple baseline, there are little to no comparisons of their effectiveness relative to each other.

## References

- [1] M. Maimaiti, Y. Liu, H. Luan, M. Sun, Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation, *Tsinghua Science and Technology* 27 (2022) 150–163.
- [2] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, G. Neubig, When and why are pre-trained word embeddings useful for neural machine translation?, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Association for Computational Linguistics, 2018, pp. 529–535.
- [3] T. W. Guntara, A. F. Aji, R. E. Prasajo, Benchmarking multidomain English-Indonesian machine translation, in: *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, European Language Resources Association, 2020, pp. 35–43.
- [4] R. Fatullayev, A. Abbasov, A. Fatullayev, Dilmanc is the 1st mt system for azerbaijani (2008) 63–64.
- [5] G. Neubig, J. Hu, Rapid adaptation of neural machine translation to new languages, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2018, pp. 875–880.
- [6] Y. Kim, Y. Gao, H. Ney, Effective cross-lingual transfer of neural machine translation models without shared vocabularies, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2019, pp. 1246–1257.
- [7] R. Sennrich, B. Haddow, 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)*, Association for Computational Linguistics, 2016, pp. 1715–1725.
- [8] T. Kudo, Subword regularization: Improving neural network translation models with multiple subword candidates, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2018, pp. 66–75.
- [9] W. Chen, B. Fazio, Morphologically-guided segmentation for translation of agglutinative low-resource languages, in: *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, Association for Machine Translation in the Americas, 2021, pp. 20–31.
- [10] A. Richburg, R. Eskander, S. Muresan, M. Carpuat, An evaluation of subword segmentation

- strategies for neural machine translation of morphologically rich languages, in: Proceedings of the The Fourth Widening Natural Language Processing Workshop, Association for Computational Linguistics, 2020, pp. 151–155.
- [11] K. Bostrom, G. Durrett, Byte pair encoding is suboptimal for language model pretraining, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 4617–4624.
  - [12] M. Mager, A. Oncevay, E. Mager, K. Kann, N. T. Vu, Bpe vs. morphological segmentation: A case study on machine translation of four polysynthetic languages, arXiv preprint arXiv:2203.08954 (2022).
  - [13] J. Sälevä, C. Lignos, The effectiveness of morphology-aware segmentation in low-resource neural machine translation, arXiv preprint arXiv:2103.11189 (2021).
  - [14] D. Ataman, M. Negri, M. Turchi, M. Federico, Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. (2017).
  - [15] C. Lignos, Learning from unseen data, in: Proceedings of the Morpho Challenge 2010 Workshop, 2010, pp. 35–38.
  - [16] W. Chen, B. Fazio, The UCF systems for the LoResMT 2021 machine translation shared task, in: Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021), Association for Machine Translation in the Americas, Virtual, 2021, pp. 129–133.
  - [17] A. K. Ojha, C.-H. Liu, K. Kann, J. Ortega, S. Shatam, T. Fransen, Findings of the LoResMT 2021 shared task on COVID and sign language for low-resource languages, in: Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021), Association for Machine Translation in the Americas, Virtual, 2021, pp. 114–123.
  - [18] R. Vázquez, Y. Scherrer, S. Virpioja, J. Tiedemann, The Helsinki submission to the AmericasNLP shared task, in: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Association for Computational Linguistics, Online, 2021, pp. 255–264.
  - [19] M. Mager, A. Oncevay, A. Ebrahimi, J. Ortega, A. Rios, A. Fan, X. Gutierrez-Vasques, L. Chiruzzo, G. Giménez-Lugo, R. Ramos, I. V. Meza Ruiz, R. Coto-Solano, A. Palmer, E. Mager-Hois, V. Chaudhary, G. Neubig, N. T. Vu, K. Kann, Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas, in: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Association for Computational Linguistics, Online, 2021, pp. 202–217.
  - [20] G. Klein, Y. Kim, Y. Deng, J. Senellart, A. Rush, OpenNMT: Open-source toolkit for neural machine translation, in: Proceedings of ACL 2017, System Demonstrations, Association for Computational Linguistics, 2017, pp. 67–72.
  - [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
  - [22] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 2685–2702.
  - [23] M. Post, A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference



- on Machine Translation: Research Papers, Association for Computational Linguistics, 2018, pp. 186–191.
- [24] K. Papineni, S. Roukos, T. Ward, 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, Association for Computational Linguistics, 2002, pp. 311–318.
- [25] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2015, pp. 392–395.
- [26] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, A. Menezes, To ship or not to ship: An extensive evaluation of automatic metrics for machine translation, in: Proceedings of the Sixth Conference on Machine Translation, Association for Computational Linguistics, 2021, pp. 483–499.
- [27] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, 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, Association for Computational Linguistics, 2021, pp. 1351–1361.
- [28] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, Transactions of the Association for Computational Linguistics 7 (2019) 597–610.
- [29] J. Ortega, R. Castro Mamani, K. Cho, Neural machine translation with a polysynthetic low resource language, Machine Translation (2021).
- [30] J. Zuters, G. Strazds, K. Immers, Semi-automatic quasi-morphological word segmentation for neural machine translation, in: A. Lupeikiene, O. Vasilecas, G. Dzemyda (Eds.), Databases and Information Systems, Springer International Publishing, 2018, pp. 289–301.
- [31] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: Open source toolkit for statistical machine translation, in: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Association for Computational Linguistics, 2007, pp. 177–180.
- [32] A. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Transactions on Information Theory 13 (1967) 260–269.
- [33] J. Tiedemann, L. Nygaard, The OPUS corpus - parallel and free: <http://logos.uio.no/opus>, in: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), 2004.
- [34] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, A. Joulin, A. Fan, CCMatrix: Mining billions of high-quality parallel sentences on the web, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, 2021, pp. 6490–6500. URL: <https://aclanthology.org/2021.acl-long.507>.
- [35] J. Tiedemann, Parallel data, tools and interfaces in opus, in: N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), 2012.

- [36] A. Abdelali, F. Guzman, H. Sajjad, 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), 2014.
- [37] N. Reimers, 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, 2020.
- [38] A. El-Kishky, A. Renduchintala, J. Cross, F. Guzmán, P. Koehn, XLEnt: Mining cross-lingual entities with lexical-semantic-phonetic word alignment, in: Preprint, 2021.