Neural Machine Translation for English–Kazakh with Morphological Segmentation and Synthetic Data

Antonio Toral† **Lukas Edman**‡ **Galiya Yeshmagambetova**‡ **Jennifer Spenader**‡
†Center for Language and Cognition, ‡Institute for Artificial Intelligence
University of Groningen
The Netherlands

a.toral.ruiz@rug.nl,{j.l.edman,g.yeshmagambetova}@student.rug.nl, j.spenader@ai.rug.nl

Abstract

This paper presents the systems submitted by the University of Groningen to the English–Kazakh language pair (both translation directions) for the WMT 2019 news translation task. We explore potential benefits from using (i) morphological segmentation (both unsupervised and rule-based), given the agglutinative nature of Kazakh, (ii) data from two additional languages (Turkish and Russian), given the scarcity of English–Kazakh data, and (iii) synthetic data, both for the source and for the target language. Our best submissions ranked second for Kazakh→English and third for English→Kazakh in terms of the BLEU automatic evaluation metric.

1 Introduction

This paper presents the neural machine translation (NMT) systems submitted by the University of Groningen to the WMT 2019 news translation task.[1] We participated in the English↔Kazakh (henceforth referred to as EN↔KK) constrained tasks.

Because of the inherent characteristics of this language pair and the current state-of-the-art of related techniques, we focused on two main research questions (RQs):

- RQ1. Does morphological segmentation help? Recent research in NMT for agglutinative languages found that morphological segmentation outperforms the most widely used segmentation technique, byte-pair encoding (BPE, using character sequence frequencies) (Sennrich et al., 2016). Rule-based segmentation improved English-to-Finnish translation (Sánchez-Cartagena and Toral, 2016) and unsupervised segmentation improved Turkish-to-English translation (Ataman et al., 2017). Because Kazakh belongs to the same language family as Turkish, the work by Ataman et al. (2017) is particularly relevant. Their training data had fewer than 300,000 sentence pairs and they trained an NMT system under the recurrent sequence-to-sequence with attention paradigm (Bahdanau et al., 2015). Our training data is considerably bigger and we use a non-recurrent attention-based system (Vaswani et al., 2017). Does the advantage of morphological segmentation over BPE also hold in our experimental setup?

- RQ2. Does the use of additional languages improve outcomes? Due to the scarcity of parallel data for EN–KK, we investigate if using data from two additional languages is useful, Russian (RU) and Turkish (TR). Even though RU is not related to either EN or KK, it seems a sensible choice due to the availability of large amounts of EN–RU and RU–KK parallel data. TR is related to KK and there are limited amounts of EN–TR data available. Does this additional data improve the performance, and is more data from an unrelated language (RU) better than less data from a related language (TR)?

The rest of the paper is organized as follows. Section 2 describes the datasets and tools used. Then Section 3 details our experiments. Finally, Section 4 outlines our conclusions and plans for future work.

2 Datasets and Tools

We preprocessed all the corpora used (training, validation and test sets) with scripts from the

[1] http://www.statmt.org/wmt19/translation-task.html

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 386–392
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

Moses toolkit (Koehn et al., 2007). The following operations were performed sequentially: punctuation normalisation, tokenisation,[2] truecasing and escaping of problematic characters. The truecaser was lexicon-based and it was trained on all the monolingual data available for each language. In addition, we removed sentence pairs where either side was empty or longer than 80 tokens from the parallel corpora . Tables 1 to 4 show the parallel datasets used for training for each translation direction after preprocessing. The corpora Kazakhtv (EN–KK) and crawl (KK–RU) were provided with sentence-level scores; we sorted their files according to these scores and a native KK speaker proficient in both EN and RU identified a threshold where alignments were roughly 90% correct. This led to discarding the bottom 27% of the data for EN–KK's Kazakhtv and the bottom 3% for KK–RU's crawl.

Corpus	Sentences (k)	Words (M) EN	KK
Kazakhtv	67.7	1.00	0.82
News-comm.	7.5	0.19	0.16
Wikititles	117.0	0.23	0.19

Table 1: Preprocessed EN–KK parallel training data.

Corpus	Sentences (k)	Words (M) EN	RU
Common crawl	871.8	20.82	19.97
News-comm.	278.2	7.17	6.86
Paracrawl	11,881.0	189.90	166.50
Yandex	997.3	24.06	22.00

Table 2: Preprocessed EN–RU parallel training data.

Corpus	Sentences (k)	Words (M) KK	RU
Crawl	4,861.5	99.34	105.16

Table 3: Preprocessed KK–RU parallel training data.

All our NMT systems are trained with Marian (Junczys-Dowmunt et al., 2018).[3] We used the `transformer` model type (Vaswani et al., 2017)

[2]Moses does not contain a tokeniser for KK. KK texts were tokenised with the RU model, as both languages are written in the cyrillic alphabet. The resulting tokenisation was inspected and validated by a KK native speaker.

[3]https://marian-nmt.github.io/

Corpus	Sentences (k)	Words (M) EN	TR
newstest2016-18	9.0	0.20	0.17
SETimes	207.4	5.12	4.61

Table 4: Preprocessed EN–TR parallel training data.

in all experiments, except for a few experiments where the training data was very limited, where we used the `s2s` model type (Bahdanau et al., 2015).

During development, we evaluated our systems on the development sets provided. We used two automatic evaluation metrics: BLEU (Papineni et al., 2002) and CHRF (Popović, 2015). CHRF is our primary evaluation metric for EN→KK, due to the fact that this metric has been shown to correlate better than BLEU with human evaluation when the target language is agglutinative (Stanojević et al., 2015). BLEU is our primary evaluation metric for KK→EN systems, as the correlations with human evaluation of BLEU and CHRF are roughly on par for EN as the target language. Prior to evaluation the MT output is detruecased and detokenized with Moses' scripts.

3 Experiments

3.1 Cyrilization and Turkish

Since KK is a low-resourced language, multilingual NMT (Johnson et al., 2017) was used. Following Neubig and Hu (2018), we have chosen TR as a helper source language, because it is related to KK (both belong to the same language family) and TR is higher-resourced than KK. However, TR uses a Latin-script alphabet, while KK uses a Cyrillic-script alphabet, which means their vocabularies do not match as they are. For this reason, we decided to transliterate TR into Cyrillic (*cyrillization*). However, some characters in KK's alphabet are not present in existing transliterators. Therefore, we created a cyrillizer that matches KK's alphabet exactly.

We trained a {KK, TR}→EN system in two steps. First, we use as training data the concatenation of the EN–KK and EN–TR corpora (Tables 1 and 4) and when the model converged, we resume training using only the EN–KK dataset. We compared models that used the original TR versus cyrillized. These models were trained with the `s2s` architecture using 32,000 joining operations in BPE and dropout of 0.05.

Training data	BLEU
EN–KK	6.61
+ EN–TR	11.15
+ cyrillizer	10.34

Table 5: BLEU scores on the development set for KK→EN using additional EN–TR data.

As it is shown in Table 5, the addition of EN–TR data proves very beneficial (absolute improvement of 4.5 BLEU points), which is not surprising since the amount of training data more than doubles (cf. Tables 1 and 4). However, cyrillising TR decreases the BLEU score by 0.8 points.

3.2 Backtranslation and Russian

Given the small amount of EN–KK parallel data (see Table 1) and the large amount of EN–RU and KK–RU datasets, we introduced RU as a pivot language, using backtranslation (Sennrich et al., 2015) to derive bigger datasets where the source side is synthetic. For KK→EN, we trained a RU→KK auxiliary system on the available KK–RU data (Table 3), and used this to translate the RU portion of the EN–RU (Table 2) data into KK, creating a synthetic EN–KK' dataset. This was then used, along the original EN–KK data (Table 1) to train the KK→EN model.

For EN→KK, we trained a RU→EN auxiliary model on the available EN–RU data, and used this model to translate the RU portion of the KK–RU data into EN, creating a synthetic EN'–KK dataset. This synthetic dataset, alongside the original EN–KK data, was then used to train the EN→KK model.

Table 6 shows the results for EN→KK and KK→EN without and with the backtranslated data. The addition of backtranslated data results in massive improvements: +17.9 CHRF points for EN→KK and +14.2 BLEU points for KK→EN. This is expected given the very small size of EN–KK data and the much larger EN–RU and KK–RU datasets. The improvements are considerably larger than those obtained with additional EN–TR data (see Table 5).

Backtranslation	EN→KK	KK→EN
No	27.75	6.61
Yes	45.67	20.17

Table 6: Performance of MT systems with and without backtranslation for EN→KK (CHRF) and KK→EN (BLEU).

3.3 Corpus Filtering and Target Synthetic Data

Since most of our training data is crawled, we applied corpus filtering to remove noisy sentence pairs. Following Artetxe and Schwenk (2018a), we removed sentences shorter than 3 words and longer than 80 words, and sentence pairs where either sentence is classified as another language using the FastText language identifier (Joulin et al., 2016a,b).[4] We also removed sentence pairs with a token overlap of 50% or higher.

We identify and remove misaligned sentence pairs (where the meanings of the source and target sentences do not match), using the `LASER` system, a 93-language BiLSTM encoder (Artetxe and Schwenk, 2018b).[5] This encodes the sentences in each side, and uses the cosine similarity between the embeddings of the two sentences as a filtering threshold (where sentences below the threshold are removed).

This filtering is applied after backtranslation (see Section 3.2). For KK→EN, we filter the EN–KK' data, i.e. the EN–RU corpora whose RU side had been translated into KK. The thresholds (determined manually, as previously mentioned in Section 2) and number of sentence pairs kept are shown in Table 7.

Corpus	Threshold	Pairs left (k)
CommonCrawl	0.7323	568.50
News Comm.	0.7314	254.79
ParaCrawl	0.8031	4056.28
Yandex	0.7220	887.76

Table 7: Cosine similarity thresholds used to filter out EN–RU corpora and resulting corpus sizes after all filtering steps are applied.

We quantify the impact on translation performance of each filtering step, cumulatively, in Table 8. Each filtering step improves the BLEU score, corroborating previous research, e.g. (Koehn et al., 2018), that has shown that noisy sentence pairs indeed cause a drop in translation performance.

[4]https://fasttext.cc/docs/en/language-identification.html
[5]https://github.com/facebookresearch/LASER

Filtering	BLEU	# sent. pairs
none	20.17	15.1
language identification	20.76	9.8
+cosine	21.60	6.9
+3-80 & overlap	22.26	5.4

Table 8: BLEU scores for KK→EN systems adding one filtering mechanism at a time. The table also shows the number of sentence pairs (millions) that make up the training data for each system.

In the opposite direction, EN→KK, we filter the EN'–KK data, i.e. the RU–KK corpora whose RU side has been translated into EN. The threshold and number of sentence pairs kept are shown in Table 9.

Corpus	Threshold	Pairs left (k)
Crawl	0.1463	4494.10

Table 9: Sentence pairs left in the EN'–KK dataset after filtering.

By manual inspection, we noticed that the biggest dataset used for EN→KK (the KK–RU crawl corpus, see Table 3) is domain-specific and rather unrelated to the domain of the test set (news). Due to this, we decided to experiment with target synthetic data by translating the EN–RU corpora, which are not domain-specific, into KK and adding a subset of the resulting EN–KK' data to our EN→KK system. We experimented with two similarity thresholds: a more conservative one (0.8) and a less conservative one (0.75). The thresholds and number of sentence pairs kept are shown in Table 10.

	Pairs left (k)	
Corpus	sim $\geq$ 0.75	sim $\geq$ 0.80
CommonCrawl	80.49	30.47
News Comm.	15.41	3.71
ParaCrawl	739.16	320.98
Yandex	83.16	31.65

Table 10: Sentence pairs left in the EN–KK' dataset after filtering using the similarity thresholds 0.75 and 0.8.

Table 11 shows the impact of adding target synthetic data on translation performance. Adding a small amount using a conservative threshold (0.8) results in an absolute improvement of 1.15 CHRF points. Adding more data using a less conservative threshold (0.75) results in a bigger improvement of 1.6 points. An even lower threshold was not tested due to time constraints.

Target synthetic data	CHRF
None	45.67
similarity>0.80	46.82
similarity>0.75	47.27

Table 11: Impact of adding target synthetic data on translation performance (CHRF) for the translation direction EN→KK

3.4 Segmentation

Data is segmented with BPE (Sennrich et al., 2016) on all the languages involved in our experiments (EN, KK, RU and TR). In addition, we perform two types of morphological segmentation on KK: unsupervised and rule-based.

Unsupervised morphological segmentation is performed with LMVR (Ataman et al., 2017),[6] a variant of Morfessor (Virpioja et al., 2013) that allows a fixed vocabulary size to be defined. LMVR was trained on the KK side of the RU–KK parallel data as well as on the monolingual KK data. We experimented using vocabulary sizes of 8, 16, 24, and 32 thousand. The trained LMVR models are used to segment the KK portion of the RU–KK data and the synthetic KK derived from the EN-RU data created with a RU-KK system (see Section 3.2).

For rule-based segmentation, `Apertium-kaz` (Washington et al., 2014) was used.[7] A transducer that provides multiple segmentation variants was set up four our purpose,[8] from these variants we decided to pick the one that segments into the smallest units, because this one, as observed by manual inspection, tends to be correct more often. Some segmentations do not correspond to the original word when joined, which we attribute to the fact that Apertium is not doing pure segmentation but also analysis. We do not pick these variants. We also observed that some words were out-of-vocabulary (OOV), i.e. not found in Apertium's transducer, so those were left unsegmented.

As can be seen in Table 12, Apertium segmenter leads to lower automatic metric scores, while BPE and LVMR are on par. This could be attributed

[6] https://github.com/d-ataman/lmvr

[7] http://wiki.apertium.org/wiki/Apertium-kaz

[8] This is a version of the regular transducer that does not delete the morpheme boundary in the morphophonological rules, and is therefore more suitable for segmentation.

to the morphological ambiguity issues described above and to the fact that some words were not segmented (OOV).

Segmentation	EN→KK	KK→EN
BPE	45.67	22.26
LVMR	45.47	22.36
Apertium	42.21	-

Table 12: Performance of MT systems using different segmentations (BPE, LVMR and Apertium) for EN→KK (CHRF) and KK→EN (BLEU). Apertium was not used for the KK→EN due to time constraints.

Besides these quantitative results, we also performed qualitative analyses of the segmentations. Table 13 shows examples of words that result in ambiguous segmentations with Apertium. Table 14 shows a KK sentence segmented with BPE and LVMR. Morphological segmentation results in a better segmentation, which has a direct impact on the quality of the resulting EN translation.

3.5 Final Submissions

We took the best performing systems from previous experiments and carried out fine-tuning by resuming training after convergence using solely the EN–KK data (i.e. without any data whose source or target is synthetic). Finally, we ran ensembles of the best performing systems (with and without fine-tuning) and chose those that perform best on the development set. Those constitute our submissions to the shared task.

For KK→EN, we consider systems segmented with BPE and with LVMR since their BLEU scores are roughly on par: 22.26 and 22.36, respectively. The fine-tuned KK→EN system with BPE segmentation reaches 23.11. We built an ensemble on four BPE-based models, the two top performing ones without fine tuning (21.9 and 22.26) and the two top performing ones with fine tuning (22.99 and 23.11). The ensemble attains 23.37. We then tried different length-penalty values for the decoder (parameter `normalize` in Marian), using 0.9 (instead of the default 0.6) we reach 23.47.

The fine-tuned KK→EN with LVMR reaches a BLEU score of 23.26, thus slightly outperforming the fine-tuned system with BPE (23.11). We also performed fine-tuning including the synthetic data but including the non-synthetic data four times (i.e. synthetic to non-synthetic ratio of 1:4). This system reaches 22.65. We built an ensemble of the two fine-tuned models. This ensemble achieves a BLEU score of 23.71, which using a length normalisation penalty of 0.9 increases to 23.84.

For EN→KK we submitted systems based on BPE segmentation only. Our best of these systems achieves 47.27 CHRF while the best LVMR-based system yields 45.27.[9] We build an ensemble made of five models: the two top performing ones using target synthetic data with threshold 0.8 (CHRF scores 46.48 and 46.79), the two top performing ones using target synthetic data with threshold 0.75 (CHRF scores 47.07 and 47.27), and the top performing fine-tuned model with threshold 0.75 (CHRF score 47.57). The ensemble attains a CHRF score of 48.43.

4 Conclusions

This paper has reported on the systems submitted by the University of Groningen to the English↔Kazakh translation directions of the news shared task at WMT 2019.

Our results show quantitative evidence that, for an agglutinative language such as Kazakh, morphological segmentation is on par with segmentation based on the frequency of character sequences (in terms of automatic evaluation metrics) and qualitative evidence that it can result in better translations due to segmenting at the right morpheme boundaries. In addition, we show that the addition of data from an additional language, be it related or not, improves the performance notably, corroborating previous results. Finally, the use of synthetic data (both for the source and target languages), filtered with a state-of-the-art system based on language-independent similarity, improved the performance of our systems further.

As for future work, we plan to work along three lines. First, related to morphological segmentation, we note that Kazakh uses vowel harmony, which should be useful to model as part of the segmentation. Second, we would like to explore the contribution of synthetic target data in further detail. Third, given the unexpected negative results of cyrillization, we plan to analyse cyrillization's effects in detail.

[9]The BPE-based system uses target synthetic data while the LVMR-based system does not. The BPE-based system without target synthetic data reaches 45.67 CHRF, thus on par with the LVMR-based system (45.27 CHRF). We did not build a LVMR-based system with target synthetic data due to time constraints.

Original word	Segmentations
осыдан	осыдан | осынан
тіркелмеген	тіркелмеген | тіркел→ген емес
құжаттардың	құжат→тар→дың | құжатта→р→дың
өнерін	өнер→ін | өн→ер→ін

Table 13: Examples of morphological ambiguity challenges faced using Apertium's segmenter. The segmentation variants shown include those that when joined do not match the original word (underlined).

Segmentation	Sentence and System output
None	Қауіптің алдын алуға жәрдемдесетін мұндай құрылғыларды көптеп дайындауға облыс әкімдігі мен Қорқыт ата атындағы Қызылорда Мемлекеттік университетінің басшылығы ұсыныс білдіріпті.
BPE	Қауіп→тің алдын алуға жәрдемде→сетін мұндай құрылғыларды көпте→п дайындауға облыс әкімдігі мен Қорқы→т ата атындағы Қызылорда Мемлекеттік уни→верси→те→тінің басшыл→ығы ұсыныс білдір→іп→ті. In addition, the regional administration and the Kyzylorda State Universum named after the Fund named after the President of the Republic of Kazakhstan are ready to provide assistance in the prevention of the threat.
LVMR	Қауіп→тің алды→н ал→уға жәрдемде→сетін мұн→дай құр→ылғы→ларды көп→теп дайын да→уға облыс әкім→дігі мен Қорқыт ата ат→ындағы Қызыл→орда Мемлеке→ттік уни→верситет→інің басшылығы ұсыныс білдір→іпті. According to the Governor's Office of the region and the leadership of the Kyzylorda State University named after the Foundation of the First President of Kazakhstan, such devices are ready to help in the prevention of the threat.
English reference	Regional Akimat and Management of Kyzylorda State University named after Korkyt ata proposed to fabricate such safety devices assisting in prevention of danger in large quantities.

Table 14: Segmentation examples of BPE and unsupervised morphological segmentation (LVMR) systems for KK→EN. Arrows represent boundaries between the morphs in which a word is split. Note that the word "университетінің" is segmented differently in both systems. The MT system with LVMR segmentation translates it correctly as *"University"*, while the MT system with BPE segmentation produces *"Universum"* because of incorrect segmentation. This word, its segmentations and its translations are underlined.

Acknowledgments

We would like to thank Jonathan Washington and Francis Tyers for setting up for us a custom segmenter for `Apertium-kaz` tailored to morphological segmentation.

References

Mikel Artetxe and Holger Schwenk. 2018a. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. *arXiv preprint arXiv:1811.01136*.

Mikel Artetxe and Holger Schwenk. 2018b. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *arXiv preprint arXiv:1812.10464*.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331–342.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 739–752, Belgium, Brussels. Association for Computational Linguistics.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

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

Maja Popović. 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. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena and Antonio Toral. 2016. Abu-MaTran at WMT 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *Proceedings of the First Conference on Machine Translation*, pages 362–370, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

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

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Jonathan Washington, Ilnar Salimzyanov, and Francis Tyers. 2014. Finite-state morphological transducers for three Kypchak languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).