Multi-Parallel Corpus of North Levantine Arabic

Mateusz Krubiński[1], Hashem Sellat[1], Shadi Saleh[1], Adam Pospíšil[2], Petr Zemánek[2], and Pavel Pecina[1]

[1]Charles University, Faculty of Mathematics and Physics
{krubinski,sellat,saleh,pecina}@ufal.mff.cuni.cz
[2]Charles University, Faculty of Arts
{adam.pospisil,petr.zemanek}@ff.cuni.cz

Abstract

Low-resource Machine Translation (MT) is characterized by the scarce availability of training data and/or standardized evaluation benchmarks. In the context of Dialectal Arabic, recent works introduced several evaluation benchmarks covering both Modern Standard Arabic (MSA) and dialects, mapping, however, mostly to a single Indo-European language – English. In this work, we introduce a multi-lingual corpus consisting of 120,600 multi-parallel sentences in English, French, German, Greek, Spanish, and MSA selected from the OpenSubtitles corpus (Lison et al., 2018), which were manually translated into the North Levantine Arabic. By conducting a series of training and fine-tuning experiments, we explore how this novel resource can contribute to the research on Arabic MT. We make the dataset publicly available at http://hdl.handle.net/11234/1-5033 for research purposes.

apc	جواز سفري هنيك مع شوية وراق
	مين عم ياكل فطايري؟
arb	جواز سفري هناك مع بعض الأوراق
	من الذي يأكل فطائري ؟
eng	My passport is there, along with some papers.
	Who's eating my dumplings?
fra	Il y a mon passeport et des papiers dedans.
	Qui mange mes dumplings ?
deu	Dort drin ist mein Pass und einige Papiere.
	Wer isst meine Klöße?
ell	κεί είναι το διαβατήριό μου και μερικά έγγραφα.
	Ποιος τρώει τα ντάμπλιν μου
spa	Dentro está mi pasaporte, además de unos papeles.
	¿Quién se come mis dumplings?

Table 1: Samples from the multi-parallel corpus introduced in this work. Translations in the Indo-European languages and MSA were obtained from the OpenSubtitles-v2018 corpus, and the ones in North Levantine Arabic (apc) were manually translated from MSA (arb).

1 Introduction

Levantine Arabic is considered one of the core units within the Arabic dialectal continuum. It can be divided into at least three dialectal regions (Al-Wer and de Jong, 2017) but the most notable division within this group lies between South Levantine (Palestinian) and North Levantine (based on the urban speech of mainly Beirut and Damascus) with clear differences between the two (Kwaik et al., 2018). At the same time, North Levantine Arabic (also called Syrian or Shami) is perceived as a clearly established linguistic unit with a positive evaluation and perception (Ghobain, 2017).

In the field of Natural Language Processing, North Levantine Arabic is, similarly to other Arabic dialects, considered a low-resource language. It is mainly used for daily speech, and written resources are very scarce. Formal texts are almost exclusively written in Modern Standard Arabic (MSA). Recently, written North Levantine Arabic started to appear in texts posted to social networks that became a useful resource of monolingual datasets for several dialects of Arabic (Abdul-Mageed et al., 2020). Parallel datasets are even scarcer.

In this paper, we introduce a novel multi-parallel corpus where North Levantine Arabic is paired with MSA and several Indo-European languages (English, French, German, Greek, and Spanish). The corpus contains roughly 1 million words on the English side. By targeting the subset of the multi-parallel OpenSubtitles-v2018 (Lison et al., 2018) dataset, we ensure that with a single round of translation, we can achieve the desired multi-lingual, multi-parallel mapping between MSA, Dialectal Arabic and several Indo-European languages. Considering that the OpenSubtitles dataset consists of lines from movie subtitles[1], it should well represent the "everyday dialogue" domain, where the Arabic

[1]https://www.opensubtitles.org

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 411–417
December 7, 2023

dialects are most commonly used.

2 Related Work

In their pioneer work, Zbib et al. (2012) introduced a parallel Levantine-English corpus of 138k sentences suitable for training MT systems. The Levantine sentences were extracted from Arabic weblogs and online user groups and translated into English. In follow-up work, Bouamor et al. (2014) translated 2,000 sentences from the Egyptian-English corpus introduced by Zbib et al. (2012) into several Arabic Dialects (including North Levantine Arabic), creating the first multi-parallel corpus of multi-dialectal Arabic. The multi-parallel aspects were further explored (e.g., Bouamor et al., 2018) and the data were compiled into standardized benchmarks (e.g., Sajjad et al., 2020; Nagoudi et al., 2023; Abdelali et al., 2023). Arab-Acquis (Habash et al., 2017) matched multi-parallel corpus of 22 European languages with human translations into MSA – dialectical aspects were not considered. The exploitation of the OpenSubtitles corpus in the context of Arabic MT was previously explored by Nagoudi et al. (2022), who used it to sample training/testing data for translation from four languages (English, French, German, and Russian) into MSA and Alhafni et al. (2022) who sampled English-MSA sentence pairs for the extended Arabic Parallel Gender Corpus (APGC v2.0).

3 Data preparation

As a first step, we filtered the OpenSubtitles-v2018 corpus by identifying lines that are available in all of the desired languages (MSA, English, French, German, Greek, and Spanish), obtaining 3,661,627 sentences. Subsequently, a number of additional filters (for convenience, we applied filters to the English side) were applied:

1. Sentences containing vulgar words (based on a hand-crafted list) were removed.
2. Sentences containing non-standard characters were removed – only punctuation marks, English alphabet letters and digits were allowed.
3. To avoid incomplete sentences, only sentences that start with a capital letter were kept.
4. Very similar sentences were discarded by lowercasing the text, removing punctuation and digits, and removing the duplicates. The goal was not to translate similar sentences like *Good morning* and *Good morning!* or *I was born in 1961* and *I was born in 1983.*

Language	ISO 639-3 code	#Words
North Levantine Arabic	apc	738,812
Modern Standard Arabic	arb	802,313
English	eng	999,193
French	fra	956,208
German	deu	940,234
Greek	ell	869,543
Spanish	spa	920,922

Table 2: Word-level statistics of the multi-parallel corpus of North Levantine Arabic introduced in this work.

5. To assure the inner variance and semantic richness of the translated text, sentences with less than two words, ones containing very rare words, and sentences with a high proportion of frequent words (frequency-based approach with a manual filtering step) were removed.

Those heuristics were necessary to both filter out low-quality sentences and to down-sample the set of translation candidates to fit within the available budget. We acknowledge that potentially valuable, semantically rich utterances that e.g., do not start with a capital letter, may have been dropped.

After those filtering steps, we ended up with 120,771 sentences. Before the translation, an additional corpus-wise filtering step was applied by removing multi-parallel lines where: English characters appear in the Arabic sentence, Arabic characters appear in the English sentence, or Arabic characters appear in a particular sentence for all of the Indo-European languages. The final size of the corpus is equal to 120,600 lines that were manually translated into the North Levantine Arabic dialect.

The translation was performed by native speakers of the dialect through a professional translation company without using any MT or CAT tool. Considering the lack of official spelling standards for Levantine, we did not provide the translators with specific orthographic guidelines (Habash et al., 2018), but rather relayed on their expertise, asking only for internal consistency. First, a sample of 1,000 sentences was translated independently from English and from MSA. No difference in translation quality was observed (assessed by authors of the paper – speakers of North Levantine Arabic). Therefore, all the remaining sentences were translated from MSA (this direction was less costly). The translation was done in batches of 5,000 sentences, and the quality of the translation was checked after each batch (again by the authors of the paper – speakers of the dialect). In order to quantitatively measure the impact of the source language, we computed the Overlap Coefficient (OC) (Bouamor et al., 2014) for the samples of 1,000 sentences that were used initially[2]. The OC value measures the percentage of lexical overlap between the vocabularies of two languages (dialects). The OC similarity between the MSA source translated into apc target equals 35.95, and the one between the (parallel) MSA and the target apc when translating from English equals 26.85. To put those numbers into context, the OC value between the 1,000 sentences in MSA and Syrian that were independently translated from Egyptian by Bouamor et al. (2014) equals 39.85. Those results indicate that the variety in the apc output may have been slightly reduced by translating from MSA. However, it should be mentioned that we compare disjoint sets of sentences, and there is not enough data to say how this affects the downstream tasks, such as MT.

Sentence samples (multi-parallel lines) are presented in Table 1, and some corpus-wise word-level statistics are presented in Table 2.

4 MT Experiments

In order to demonstrate the validity of the corpus, we conducted a number of MT experiments and evaluations.

Baselines and Metrics We report the performance of two well-established baselines: a multilingual NLLB model (Costa-jussà et al., 2022), using the `facebook/nllb-200-distilled-600M` variant (600M parameters) from the Transformers (Wolf et al., 2020) package, and uni-directional models (depending on the language pair, between 76M and 240M parameters) provided by the Helsinki-NLP group (Tiedemann, 2020). To indicate to what extent MSA can be used when the dialectal system is not available, we translate into both arb (e.g., Opus_{arb}) and apc, always using the apc files as reference. We measure the output quality by reporting the surface-level chrF++[3] metric (Popović, 2015), and the trainable, estimator-based COMET[4] metric (Rei et al., 2020).

Testing data In Table 3, we report performance on the test split of FLORES-200 (Costa-jussà et al., 2022), which consists of professional translation of sentences sampled from the English Wikipedia. In Table 4, we report on the subset[5] of MADAR (Bouamor et al., 2018), which was created by translating sentences from the Basic Traveling Expression Corpus (Takezawa et al., 2007) into several country- and city-level Arabic dialects. Since the original English and French versions of the corpus are not directly available[6], we use only the English side, as provided by the AraBench (Sajjad et al., 2020) benchmark. We report only on the test-sets corresponding to Damascus and Aleppo, as we were unable to directly match the Beirut one from MADAR to the English file in AraBench.

North Levantine Corpus In order to demonstrate the importance of pre-training, we train (Base_{ML}) a multi-lingual Transformer (Vaswani et al., 2017) model from scratch, training with the default `transformer-big` configuration (200M parameters) from the Marian toolkit (Junczys-Dowmunt et al., 2018) on the multi-parallel corpus introduced in this work. We use the source-tagging approach (Johnson et al., 2017), training on all (84) available directions, with an early stopping applied if chrF++ on FLORES-200 dev-set ceases to improve for 10 consecutive evaluations.

Furthermore, we use it to fine-tune both Opus (Opus_{FT}) and NLLB (NLLB_{FT}) models. For uni-directional Opus models, we use only mono-directional data (e.g., apc-ell) and the recommended[7] parameters. We fine-tune the NLLB model on the apc-centric data (i.e., on all of the available directions with apc as source and target) using AdamW (Loshchilov and Hutter, 2019) optimizer with a constant learning rate of 1e-5, obtaining the best results after a single epoch of fine-tuning.

5 Results

Automatic metrics The Base_{ML} system trained from scratch achieves the lowest scores on both test-sets. On average, the larger, multi-lingual NLLB model achieves better scores than the Opus models. Translating into arb gives consistently higher scores for sentences from the FLORES-200 test-set, but lower ones for sentences from MADAR. We attribute this to the vastly different nature of those test-sets. Sentences in FLORES-200 are long, with

[2] We have normalized and tokenized the sentences with the CAMeL Tools (Obeid et al., 2020) package.

[3] `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.3.1`

[4] Model signature: `Unbabel/wmt22-comet-da`

[5] Lines marked as `corpus-6-test-corpus-26-test`

[6] `https://camel.abudhabi.nyu.edu/madar-parallel-corpus`

[7] `https://github.com/Helsinki-NLP/OPUS-MT-train/blob/master/finetune`

...→apc	arb		eng		fra		deu		ell		spa	
	ChrF	COMET	ChrF	COMET	ChrF	COMET	ChrF	COMET	ChrF	COMET	ChrF	COMET
Opus_{arb}	-	-	**51.47**	**.836**	**38.54**	**.800**	**42.98**	**.799**	30.87	**.745**	35.87	**.791**
Opus_{apc}	-	-	**50.55**	**.825**	38.28	.795	37.54	.749	-	-	34.77	.777
Opus_{FT}	-	-	48.48	.786	35.70	.725	**39.24**	.730	31.47	.698	33.56	.722
Base_{ML}	13.17	.449	12.55	.431	12.61	.425	12.42	.414	12.45	.437	12.44	.427
NLLB_{arb}	<u>**47.72**</u>	<u>**.882**</u>	45.38	.824	**39.05**	**.800**	38.68	**.787**	**35.79**	**.784**	**36.21**	**.794**
NLLB_{apc}	44.12	**.832**	43.43	.795	37.03	.759	36.22	.735	33.63	.743	34.47	.756
NLLB_{FT}	**49.60**	.823	44.50	.773	38.11	.737	36.96	.718	**35.46**	.731	**36.09**	.739
apc→...												
Opus	-	-	58.13	.803	47.43	.705	46.33	.736	37.28	.750	41.42	.718
Opus_{FT}	-	-	**60.53**	**.837**	47.37	.730	**48.70**	**.769**	37.24	.773	41.66	.749
Base_{ML}	12.08	.425	16.92	.427	15.26	.357	15.80	.325	13.44	.420	16.24	.391
NLLB	**50.16**	**.854**	**59.97**	**.833**	**53.15**	**.783**	**47.19**	**.757**	**41.25**	**.818**	**44.96**	**.785**
NLLB_{FT}	**50.51**	**.854**	58.19	.831	**50.99**	**.777**	45.39	.749	**39.96**	**.811**	**44.26**	**.781**

Table 3: Evaluation results on the FLORES-200 test-set. The two highest-scoring systems in each column are bolded independently for apc source/target. Underlined numbers correspond to a copy-source system. The Greek Opus model does not support dialectal Arabic in the output.

eng→apc	Damascus		Aleppo	
	ChrF	COMET	ChrF	COMET
Opus_{arb}	26.09	**.770**	25.64	**.761**
Opus_{apc}	26.32	.757	25.71	.748
Opus_{FT}	**38.50**	.754	**40.57**	**.765**
Base_{ML}	19.01	.599	18.78	.599
NLLB_{arb}	24.58	**.761**	24.68	.753
NLLB_{apc}	33.04	.738	33.25	.739
NLLB_{FT}	**37.77**	.756	**37.30**	.756
apc→eng				
Opus	38.53	.689	39.08	.675
Opus_{FT}	51.09	.795	51.27	.780
Base_{ML}	29.08	.600	26.92	.576
NLLB	**56.21**	**.823**	**57.11**	**.815**
NLLB_{FT}	**52.91**	**.821**	**54.74**	**.804**

Table 4: Evaluation results on the subset of MADAR test-set. The two highest-scoring systems in each column are bolded independently for apc source/target.

MADAR	arb	apc	apc FT
NLLB	$2.23 \pm .30$	$2.03 \pm .08$	$\mathbf{1.54 \pm .21}$
Opus	$2.07 \pm .10$	$2.01 \pm .21$	$\mathbf{1.25 \pm .25}$
FLORES			
NLLB	$2.07 \pm .51$	$2.14 \pm .23$	$\mathbf{1.72 \pm .37}$
Opus	$1.98 \pm .19$	$2.02 \pm .24$	$\mathbf{1.57 \pm .34}$

Table 5: Results of the human evaluation. Scores indicate an average rank assigned to a sentence (lower = better). The lowest-ranked output in each row is bolded.

a high proportion of named entities (e.g., *Throughout 1960s, Brzezinski worked for John F. Kennedy as his advisor and then the Lyndon B. Johnson administration.*), while the ones in MADAR are short and simple (e.g., *Here is my passport.* or *Does that include tax?*).

The effects of fine-tuning on the corpus that we introduce highlight the difficulties of low-resource MT. On the MADAR test-set, coming from a similar domain as the resource introduced in this work, significant improvements can be observed when translating into apc – both for Opus (26.32→38.50) and NLLB (33.04→37.77) models. Similar behavior can be observed for the Opus model when translating into English (38.53→51.09). However, that is not the case for the NLLB model. It is possible that a comparable amount of dialectal Arabic (mixed with MSA) has already been seen on the source side during training, and more sophisticated fine-tuning schemas are required. On the FLORES-200 test-set (different domain), minor improvements can be observed for the NLLB model (on average, +1.97 ChrF when translating into apc), with inconsistent results for the Opus models (37.54→39.24 when translating from deu but 34.77→33.56 when translating from spa).

Human evaluation In order to verify the observations based on automatic metrics, a round of human evaluation was conducted. Two apc speakers were tasked with ranking outputs (translations of the same English sentence) from three systems: one translating into arb, one into apc, and the third one obtained by fine-tuning on the corpus introduced in this work (apc FT), in the context of the English source. The ranking procedure was done independently for both test-sets and both baseline models: NLLB and Opus – our intention was not to compare different MT models but to investigate subtle differences in the translation process. Each annotator scored 200 sentences sampled from FLORES-200 (100 unique and 100 from a control batch used to compute agreement) and 140 sampled from MADAR (60 unique and 80 common). Sentences and model outputs were shuffled to avoid positional bias. Annotators were asked to consider both fluency and adequacy of translations but to prefer the dialectal output. They were not explicitly informed that one of the translations was into `arb`, giving them the opportunity to rank it higher if the translation was perceived as more natural in the context, e.g., when translating scientific terms or if the dialectal output was ungrammatical.

The cumulative results are summarized in Table 5. In every case, on average, the output of the fine-tuned model is considered the best. On the MADAR test-set, with simple sentences, `apc` output is preferred, while on the FLORES-200 one, with long and complex ones, `arb` output is preferred. The raw inter-annotator agreement (the proportion of times both annotators ranked the same sentence equally) equals 0.52, and Cohen's κ, computed[8] with the WMT formulation for rank-based evaluation (Bojar et al., 2016), equals 0.39, indicating (Landis and Koch, 1977) a "fair/moderate" agreement.

6 Conclusions

In this work, a novel, multi-parallel corpus of North Levantine Arabic, based on the OpenSubtitles-v2018 dataset, is introduced. By fine-tuning well-established baseline MT models, we show that the dialectal aspects of language are partially orthogonal to the domain-specific properties – a dialect-specific model fine-tuned on data from a particular domain may perform worse than a more generic model if a domain shift occurs during testing. However, human evaluation confirms that the dialect-specific aspects of the output are still ranked higher and more appreciated by the final users of the MT system.

7 Acknowledgements

This work was supported by the European Commission via its H2020 Program (contract no. 870930) and CELSA (project no. 19/018) and has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (project no. LM2018101). We thank the anonymous reviewers for their valuable feedback.

[8] `https://github.com/cfedermann/wmt16/blob/master/scripts/compute_agreement_scores.py`

Limitations

Multi-parallel alignment. While a number of steps were taken to ensure the quality of the translations provided, it is possible that the multi-parallel alignments may not be perfect with languages different from the one that was used as a source. The OpenSubtitles corpus that we sub-sample from was created semi-automatically.

Multi- vs Uni-directional fine-tuning. When fine-tuning the NLLB model, we use data from all directions – with `apc` as the source and as the target. One could also consider uni-directional fine-tuning, e.g., only on the `spa-apc` direction (we explore this variant with the Opus models).

Fine-tuning on mixed data. In our experiments, we use only the corpus introduced in this work for fine-tuning. Better results could be potentially obtained by using mixed data – either with other dialectal datasets or with samples from the high-resource `arb`.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2023. Benchmarking arabic ai with large language models.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Enam Al-Wer and Rudolf de Jong. 2017. *Dialects of Arabic*, chapter 32. John Wiley & Sons, Ltd.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. The Arabic parallel gender corpus 2.0: Extensions and analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France. European Language Resources Association.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie

Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Elham Abdullah Ghobain. 2017. Dubbing melodramas in the arab world; between the standard language and colloquial dialects. *The Arabic Language and Literature*, 2:49.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nizar Habash, Nasser Zalmout, Dima Taji, Hieu Hoang, and Maverick Alzate. 2017. A parallel corpus for evaluating machine translation between Arabic and European languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 235–241, Valencia, Spain. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

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.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A lexical distance study of arabic dialects. *Procedia Computer Science*, 142:2–13. Arabic Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

El Moatez Billah Nagoudi, Ahmed El-Shangiti, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for arabic nlg.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.