Fine-Grained Language Relatedness for Zero-Shot Silesian-English Translation

Edoardo Signoroni

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
e.signoroni@mail.muni.cz

Abstract. When parallel corpora are not available to train or fine-tune Machine Translation (MT) systems, one solution is to use data from a related language, and operate in a zero-shot setting. We explore the behaviour and performance of two pre-trained Large Language Models (LLMs) for zero-shot Silesian-English translation, by fine-tuning them on increasingly related languages. Our experiment shows that using data from related languages generally improves the zero-shot translation performance for our language pair, but the optimal fine-grained choice inside the Slavic language family is non-trivial and depends on the model characteristics.

Introduction

To date, out of the 7000+[1] languages of the Earth, less than 2%[2] is covered by the machine translation systems available to the public.

Roughly half of the existing languages do not have any data that can be employed in machine translation [9]. In such cases, one strategy one could employ is to rely on related data and models to operate a zero-shot translation of the resource-scarce language pair. Previous work shows that training or transfer learning among related languages improves the performance for the low-resource pair.

However, most of this work focused on training and then fine-tuning systems from scratch. Language relatedness is also looked at horizontally, usually considering high-level language families. In this paper, we explore relatedness with increasingly fine-grained degree of relatedness with a study inside the Slavic language family, focussing on Silesian-English zero-shot translation.

We fine-tune pretrained multilingual T5 [21] variants, the subword-based mT5 [29] and the byte-level ByT5 [28], enabling for a comparison between the two processing methodologies. We evaluate the output translations with two automated metrics, ChrF++ [20] and COMET [22].

[1] Ethnologue (https://www.ethnologue.com/) lists 7168 languages, of which 3072 are endangered.

[2] As of November 2023, Google Translate supports 133 languages.

We find that using data from related languages generally improves the zero-shot translation performance for our language pair, with the greater improvement between the unrelated language and one from the same high-level language family. The results, however, also show that the behaviour of the models at a finer-grained scale is more complex and depends on the model characteristics.

1 Related Work

1.1 Language Relatedness

Previous studies have investigated language relatedness for transfer learning and MT, with most of the work focussing on training and fine-tuning multilingual models based on the Transformer [27] or Recurrent Neural Networks.

Zoph et al. (2016) [30] show that a French-English parent is better than a German-English one to initialize a Spanish-English model when trying to improve translation quality. Spanish is linguistically closer to French than German.

Dabre et al. (2017) [3] build on the work of Zoph et al. (2016) and expand the experiment in a multilingual setting. They show that transfer learning from an X-Y language pair to a Z-Y language pair has a maximum impact when the second pair is resource scarce and X and Z are in the same or similar language family.

Nguyen and Chiang (2018) [17] improve on the method from Zoph et al. and focuses on exploiting the shared lexicon of related low-resource languages. Their work is made more efficient by Kocmi and Bojar (2018) [11].

Lakew et al. (2019) [14] explore the adaptation of multilingual neural MT models to unseen languages. They find that using language model perplexity as a relatedness proxy to select the most relevant data to the test language improves translation, even in zero-shot situations.

Khatri et al. (2021) [10] focus on Indic languages and show that training a multilingual system on related languages improves the translation performance for their setting.

Edman et al. (2021) [4] applied a novel method for initializing the vocabulary of an unseen low-resource language from a related one, which resulted in an increased translation performance.

1.2 *T5 models

Raffel et al. (2019) [21] describe the "Text-to-Text Transfer Transformer" (T5), a multitask encoder-decoder LLM based on the Transformer architecture. T5 is trained on the "Colossal Clean Crawled Corpus" (C4), a heuristically-cleaned version of the Common Crawl web dump containing about 750GB of English text. T5 uses a unified "text-to-text" format for all text-based NLP problems.

Xue et al. (2021) [29] present mT5, a multilingual variant of T5 trained on a Common Crawl-based dataset covering 101 languages, called mC4. mT5 is a subword-based model, with a vocabulary of 250k SentencePiece [13] tokens. The authors focus on zero-shot generation with the aim of preventing accidental translation when evaluating generative multilingual LLMs in a zero-shot setting.

Both mT5 and its byte-level variant ByT5 have been released in five model sizes: *Small* (300M parameters), *Base* (580M), *Large* (1.2B), *XL* (3.7B), and *XXL* (13B).

Xue et al. (2022) [28] details ByT5, a token-free version of mT5 which works directly on UTF-8 byte sequences, resulting in a vocabulary of 256 possible values, thus reducing the parameters allocated to the vocabulary from 85% to 0.3% for the *Small* model. Therefore, ByT5 can process text in any language, it is more robust to noise, performs better at spelling-sensitive tasks, and does not require complex preprocessing pipelines. It is competitive with subword baselines with 4x less training text, but it has greater training and inference times, due to the increased length of byte sequences.

1.3 Evaluation Metrics

Machine translation is commonly evaluated by comparing the generated text with a reference translation through automated metrics.

ChrF++ [20] is a lexical overlap-based metric includes word bigrams to the character n-gram F-score metric proposed by Popović (2015) [19]. It calculates word and character level F-scores and then averages them together. This metric correlates stronger with human judgements than previous lexical-based metrics, such as BLEU [18] by better matching morphological variants of words.

COMET [22] (Crosslingual Optimized Metric for Evaluation of Translation) is a learned metric originally fine-tuned to estimate a Direct Assessment (DA) score [7] for a given translation by comparing it to source and reference embeddings. It was trained on top of XLM-R-large [2] on a corpus of human judgements of automated translations, both as DA or following the Multidimensional Quality Metric framework [15].

1.4 Parallel Corpora

The MaCoCu project is aimed at building monolingual and parallel corpora for under-resourced European languages by crawling large amounts of textual data from top-level domains of the Internet, and then applying a curation and enrichment pipeline [1]. It covers 17 languages, 8 (Bosnian, Bulgarian, Croatian, Macedonian, Montenegrin, Serbian, Slovene, Ukrainian) are Slavic.

The WikiMatrix [24] project extracted 135 million parallel sentences for 1620 different language pairs using massive multilingual sentence embeddings to automatically extract parallel sentences from the content of Wikipedia articles in 96 languages, including several dialects and low-resource languages. We used the Polish-English section of this corpus.

CzEng 2.0 [12] is an updated version of the CzEng parallel corpus containing 188 million parallel Czech-English sentences spanning multiple sources and domains.

Goyal et al. (2022) [6] release the Flores evaluation benchmark, consisting of 3001 sentences extracted from English Wikipedia translated in 200 languages by professional translators. This enables better assessment of model quality on low-resource languages. We use the Silesian portion as our zero-shot source.

Language	ISO Code	Group	Script	Classification
Silesian	szl	West Slavic, Lechitic, Polish-Silesian	Latin	-
Polish	pol	West Slavic, Lechitic, Polish-Silesian	Latin	4
Czech	ces	West Slavic, Lechitic, Czech-Slovak	Latin	3
Croatian	hrv	South Slavic, Western South Slavic	Latin	2
Serbian	srp	South Slavic, Western South Slavic	Cyrillic	1
Ukrainian	ukr	East Slavic, Ukrainian-Rusyn	Cyrillic	1
Maltese	mlt	Afro-Asiatic, Semitic, Arabic, ...	Latin	0

Table 1. Summary of the language selection for the experiment. The last column gives the relatedness degree we assigned to each language, from 0 (completely unrelated) to 4 (closely related). These roughly correspond to the taxonomy of the language with respect to Silesian. Croatian, Serbian, and Ukrainian are on the same level of the taxonomy, but we assigned a higher score to Croatian by virtue of it sharing the same script with Silesian.

2 Methodology

2.1 Languages, Models, and Metrics Selection

The first step of the experiment consisted in finding a proper dataset that allowed for an as clean as possible comparison. The Flores benchmark dataset features Silesian, a West Slavic language of the Lechitic subgroup, mostly spoken in Upper Silesia, Poland. Joshi et al. (2020) [9] lists Silesian as a low-resource language in terms of availability of data and research. [3]

To find data for related Slavic languages, we turned to the MaCoCu project, which evaluation [4] shows it having a significantly better quality than other web-crawled parallel corpora. Following the taxonomy in Glottolog [8], we selected 6 Slavic languages from the corpus, summarized in Table 1. The furthest removed from Silesian are Croatian, Serbian (South Slavic), and Ukrainian (East Slavic). The latter two, being written in Cyrillic script, do not even share the same writing system of Silesian. As our control language, we chose Maltese, a Semitic language also part of the MaCoCu selection.

Since the MaCoCu corpus does not cover any West Slavic language, we had to look elsewhere for languages closer to Silesian. We decided to use Czech as a West Slavic language not belonging to the Lechitic subgroup. We chose to use the CzEng 2.0 parallel corpus. As the closest language to Silesian, we selected Polish, part of the same Polish-Silesian branch of the Lechitic subgroup. The Polish data is taken from WikiMatrix.

With regard to the pre-trained models, we chose mT5-small and ByT5-small. Their similarity in training and architecture allows for a clearer comparison between subword and character-level models. Both were pretrained on the mC4 multilingual corpus, which contains data for some of the languages in our experiments and other Slavic languages in general.

[3] However, the OPUS repository [26] lists some Silesian-English parallel data available, with the NLLB [25] one consisting of 1.8 million sentences.

[4] https://macocu.eu/static/media/second-report.453a82100b1ec3647012.pdf (Retrieved on Nov 4, 2023)

Fine-Tuning	ChrF++		COMET	
Language	ByT5	mT5	ByT5	mT5
4_pol_Latn	**39.6**	29.19	**0.56**	0.45
3_ces_Latn	34.87	28.09	0.48	*0.42*
2_hrv_Latn	33.22	28.92	0.47	**0.47**
1_ukr_Cyrl	34.12	29.09	0.5	0.44
1_srp_Cyrl	33.73	**29.36**	0.5	0.46
0_mlt_Latn	*25.43*	*24.77*	*0.4*	0.44

Table 2. ChrF++ and COMET scores for each system. The best system is given in **bold** and the worst in *italic*.

Studies such as the one by Mathur et al. (2020) [16] argue for the retirement of BLEU in favour of ChrF++. Moreover, Sai B. et al. (2023) [23] finds that ChrF++ performs the best among overlap metrics for a selection of Indic languages.

However, both the aforementioned studies and the results of recent WMT Metrics shared tasks [5] demonstrate that learned neural metrics are the most optimal, as they better correlate with human judgements. Among these, COMET is the current state-of-the-art, and is widely employed in machine translation studies.

2.2 Experimental Setup

We first fine-tune translation models from each related language into English on a random sample of 250k sentence pairs. Using the HuggingFace framework, we train for a maximum of 4000 steps with a learning rate of 1e-4 and batches of 5000 tokens, with early-stopping according to the validation performance on the "dev" split of Flores-200.

To evaluate zero-shot performance, we generate English translations for the Silesian "devtest" section of Flores-200 using the fine-tuned model for each language. We then score the output with ChrF++ and COMET, using the implementations provided by HuggingFace.

3 Results

Figure 1 and Table 2 report both the Chrf++ and COMET scores for the zero-shot Silesian-English translation. From the plots, it is clear that the two models behave quite differently, with ByT5 models almost always performing better than the mT5 ones.

For ByT5, the trend is similar across the two metrics: as expected, the lowest score is for the system trained on Maltese with 25.43 ChrF++ and 0.4 COMET, while the best performance is achieved by the Polish model with 39.6 ChrF++ and 0.56 COMET. Between the two extremes, however, the trend becomes murkier. The performance for the first two related languages, Serbian and Ukrainian is similar, at around 34 ChrF++ and 0.5 COMET, and considerably better than the unrelated language. However, as we move to Croatian, the scores dip to 30.77 ChrF++ and 0.47 COMET. With Czech, the performance increases again to 34.87 ChrF++ and 0.48 COMET. The scores for Croatian and Czech

also highlight that this trend seems to be more marked for COMET scores, with the ChrF++ curve being still almost flat.

The behaviour of mT5 is even more complex. According to ChrF++, the only significative jump in performance is between Maltese at 24.77 points and all the Slavic languages, which scores lie around 28/29 points. Interestingly, the best system is the Serbian one, but just for a meagre 0.17 ChrF++. However, the scores for all the Slavic mT5 systems are so close together that no observation apart from that using a Slavic language instead of an unrelated one leads to better zero-shot performance on Silesian.

As with ByT5, the COMET plot for mT5 systems appears to be more varied. Two main points come up: first, the Maltese system performs on-par or even better than some other systems trained on related Slavic languages. It is just 0.1 COMET away from the Polish system, which sits at 0.45 points, and beats the Czech system by 0.2 COMET. Second, the best performance is obtained with Croatian fine-tuning, at 0.47 COMET.

Language	**Tokens (in Billions)**	**mC4 %**
pol	130	2.15
ces	63	1.72
ukr	41	1.51
srp	4.5	0.72
hrv	0	0
all_slavic	1005.09	15.2
mlt	5.2	0.64

Table 3. Number of tokens (in billions) and representation (as percentage of the training corpus mC4) for Slavic languages and Maltese in ByT5 and mT5. The languages are given in ISO-639-3 codes. Croatian (hrv) is not in mC4.

Table 3 gives the amount of pretraining data in the mC4 corpus for the relevant languages in our experiment. The amount of seen data for a given language does not seem to strongly impact the performance on zero-shot translation from Silesian. While it is true that Polish is by far the most represented language of the sample in pretraining, it is also the case that the model fine-tuned on Croatian, which is not present in the mC4 corpus, does not perform significantly worse than the others.

The scores for the fine-tuned systems when translating from the language of training into English is given in Table 4. The quality of the fine-tuned systems

Language	**mlt**		**srp**		**ukr**		**hrv**		**ces**		**pol**	
Model	**ByT5**	**mT5**	**ByT5**	**mT5**	**ByT5**	**mT5**	**ByT5**	**mT5**	**ByT5**	**mT5**	**ByT5**	**mT5**
ChrF++	57.23	43.54	49.34	43.33	46.74	42.64	47.8	38.74	47.85	42.7	41.6	36.29
COMET	0.67	0.57	0.72	0.69	0.72	0.7	0.73	0.64	0.73	0.7	0.7	0.66

Table 4. ChrF++ and COMET scores for the fine-tuned systems when translating from the language of training into English.

on seen source translation similarly does not appear to affect zero-shot translation. While, according to COMET, the performance is roughly at the same level for all systems, looking at ChrF++ gives another picture. As expected, the Maltese ByT5-system seems unable to overcome the typological distance when translating from Silesian, even despite its greater score. The much worse, at least according to ChrF++, Polish ByT5 system is much better for Silesian, losing just 2 ChrF++ points in the zero-shot scenario. This closeness in performance is most probably due to the high degree of relatedness between Polish and Silesian.

Overall, these results seem to indicate that language relatedness plays a part in the zero-shot translation from Silesian to English. Especially for ByT5, it is clear that fine-tuning the system for a related Slavic language improves the translation. While the closest language, Polish, performs the best for ByT5, the same cannot be said for mT5. Moreover, the impact of relatedness on a more fine-grained scale has to be further clarified, with performance fluctuating among the Slavic languages apart from Polish and with the subword model in particular.

4 Conclusions

In this paper, we described our experiment on the impact of related language fine-tuning of multilingual pretrained models for Silesian-English zero-shot translation. We compared the performance of subword-based mT5 and byte-based ByT5 models fine-tuned on a fine-grained selection of increasing related Slavic languages. Using related language data for fine-tuning seems to be beneficial in most of the cases, and while there seems to be an overall upward trend for byte models, the impact of relatedness at a finer-grained scale is still to be clarified. The representation of the fine-tuning language in the pre-trained model and the performance of the fine-tuned system translating from the seen source does not seem to play a part in our zero-shot scenario.

Limitations and Future work

This work covers just one narrow case of source-side zero-shot translation. The experiment may be expanded to other language pairs and model sizes, since the behaviour of the smaller models may differ from the larger ones.

While we tried to use comparable data from only one source for fine-tuning, for at least two languages, Czech and Polish, this was not completely possible, as they were not covered by the MaCoCu project. This can be an issue, especially with the Polish data, that are exclusively from the same domain as the Flores-200 test set. The Polish systems do not perform consistently better than the others, and thus domain similarity could play a smaller role than anticipated.

Acknowledgments

The work described herein has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

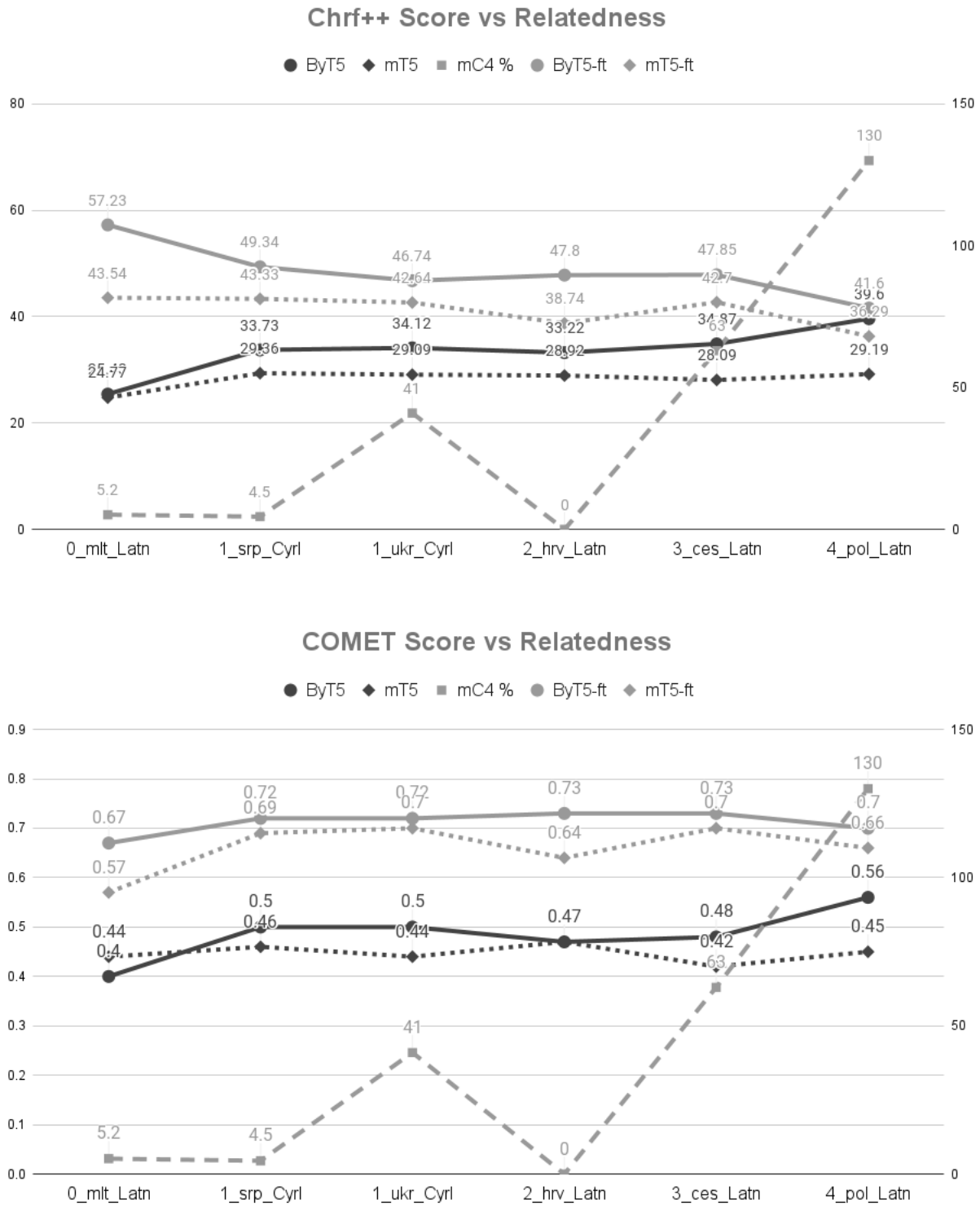

Fig. 1. Plots of ChrF++ and COMET scores for mT5 (dotted line) and ByT5 (full line) models, in order of language relatedness. The left Y-axis reports the scores, while the X-axis gives the fine-tuning language, following the Flores naming conventions. The right Y-axis shows the amount of tokens (in billions) present in the mC4 corpus for each language. The brighter lines represent the score of the fine-tuned system when translating from a seen source.

References

1. Bañón, M., Esplà-Gomis, M., Forcada, M.L., García-Romero, C., Kuzman, T., Ljubešić, N., van Noord, R., Sempere, L.P., Ramírez-Sánchez, G., Rupnik, P., Suchomel, V., Toral, A., van der Werff, T., Zaragoza, J.: MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In: Proceedings of the 23rd Annual Conference of the European Association for Machine Translation. pp. 303–304. European Association for Machine Translation, Ghent, Belgium (Jun 2022), https://aclanthology.org/2022.eamt-1.41
2. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
3. Dabre, R., Nakagawa, T., Kazawa, H.: An empirical study of language relatedness for transfer learning in neural machine translation. In: Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation. pp. 282–286. The National University (Phillippines) (Nov 2017), https://aclanthology.org/Y17-1038
4. Edman, L., Üstün, A., Toral, A., van Noord, G.: Unsupervised translation of German–Lower Sorbian: Exploring training and novel transfer methods on a low-resource language. In: Proceedings of the Sixth Conference on Machine Translation. pp. 982–988. Association for Computational Linguistics, Online (Nov 2021), https://aclanthology.org/2021.wmt-1.104
5. Freitag, M., Rei, R., Mathur, N., Lo, C.k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., Martins, A.F.T.: Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 46–68. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), https://aclanthology.org/2022.wmt-1.2
6. Goyal, N., Gao, C., Chaudhary, V., Chen, P.J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., Fan, A.: The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics **10**, 522–538 (2022). https://doi.org/10.1162/tacl$_{a}$0474, $https://aclanthology.org/2022.tacl-1.30$
7. Graham, Y., Baldwin, T., Moffat, A., Zobel, J.: Continuous measurement scales in human evaluation of machine translation. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. pp. 33–41. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013), https://aclanthology.org/W13-2305
8. Hammarström, H., Forkel, R., Haspelmath, M., Bank, S.: glottolog/glottolog: Glottolog database 4.8 (Jul 2023). https://doi.org/10.5281/ZENODO.8131084, https://zenodo.org/record/8131084
9. Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6282–6293. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.560, https://aclanthology.org/2020.acl-main.560

10. Khatri, J., Saini, N., Bhattacharyya, P.: Language relatedness and lexical closeness can help improve multilingual NMT: IITBombay@MultiIndicNMT WAT2021. In: Proceedings of the 8th Workshop on Asian Translation (WAT2021). pp. 217–223. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.wat-1.26, https://aclanthology.org/2021.wat-1.26
11. Kocmi, T., Bojar, O.: Trivial transfer learning for low-resource neural machine translation. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 244–252. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-6325, https://aclanthology.org/W18-6325
12. Kocmi, T., Popel, M., Bojar, O.: Announcing czeng 2.0 parallel corpus with over 2 gigawords (2020)
13. Kudo, T., Richardson, J.: 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. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2012, https://aclanthology.org/D18-2012
14. Lakew, S.M., Karakanta, A., Federico, M., Negri, M., Turchi, M.: Adapting multilingual neural machine translation to unseen languages (2019)
15. Lommel, A., Uszkoreit, H., Burchardt, A.: Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. Tradumàtica tecnologies de la traducció (12), 455–463 (Dec 2014). https://doi.org/10.5565/rev/tradumatica.77
16. Mathur, N., Baldwin, T., Cohn, T.: Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4984–4997. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.448, https://aclanthology.org/2020.acl-main.448
17. Nguyen, T.Q., Chiang, D.: Transfer learning across low-resource, related languages for neural machine translation (2017)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040
19. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/W15-3049, https://aclanthology.org/W15-3049
20. Popović, M.: chrf++: words helping character n-grams. In: Proceedings of the Second Conference on Machine Translation. p. 612–618. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/W17-4770, http://aclweb.org/anthology/W17-4770
21. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2019)
22. Rei, R., Stewart, C., Farinha, A.C., Lavie, A.: COMET: A neural framework for MT evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2685–2702. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.213, https://aclanthology.org/2020.emnlp-main.213

23. Sai B, A., Dixit, T., Nagarajan, V., Kunchukuttan, A., Kumar, P., Khapra, M.M., Dabre, R.: IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14210–14228. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.795, https://aclanthology.org/2023.acl-long.795
24. Schwenk, H., Chaudhary, V., Sun, S., Gong, H., Guzmán, F.: 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. pp. 1351–1361. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.115, https://aclanthology.org/2021.eacl-main.115
25. Team, N.: No language left behind: Scaling human-centered machine translation (2022)
26. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 2214–2218. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/463$_{P}aper.pdf$
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
28. Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., Raffel, C.: ByT5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics **10**, 291–306 (2022). https://doi.org/10.1162/tacl$_{a}$0461, $https://aclanthology.org/2022.tacl-1.17$
29. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: 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. pp. 483–498. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.41, https://aclanthology.org/2021.naacl-main.41
30. Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1568–1575. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1163, https://aclanthology.org/D16-1163