Beyond Literal Token Overlap: Token Alignability for Multilinguality

Katharina Hämmerl[1,2], **Tomasz Limisiewicz**[3], **Jindřich Libovický**[3], **Alexander Fraser**[4,2]

[1]Centre for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning
[3]Faculty of Mathematics and Physics, Charles University, Czech Republic
[4]Technical University of Munich, Germany

Correspondence: haemmerl [at] cis [dot] lmu [dot] de

Abstract

Previous work has considered token overlap, or even similarity of token distributions, as predictors for multilinguality and cross-lingual knowledge transfer in language models. However, these very literal metrics assign large distances to language pairs with different scripts, which can nevertheless show good cross-linguality. This limits the explanatory strength of token overlap for knowledge transfer between language pairs that use distinct scripts or follow different orthographic conventions. In this paper, we propose *subword token alignability* as a new way to understand the impact and quality of multilingual tokenisation. In particular, this metric predicts multilinguality much better when scripts are disparate and the overlap of literal tokens is low. We analyse this metric in the context of both encoder and decoder models, look at data size as a potential distractor, and discuss how this insight may be applied to multilingual tokenisation in future work. We recommend our subword token alignability metric for identifying optimal language pairs for cross-lingual transfer, as well as to guide the construction of better multilingual tokenisers in the future. We publish our code and reproducibility details[1].

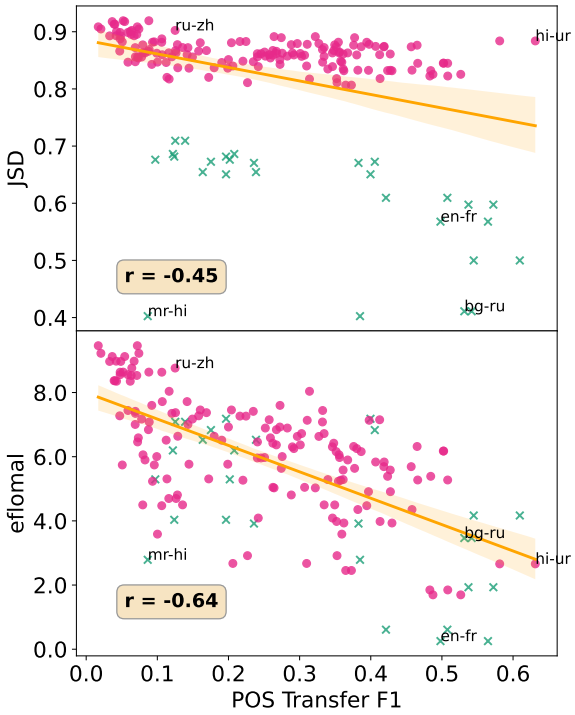

Figure 1: Eflomal score (bottom), a measure of token alignability, predicts downstream transfer performance better than the previous metric of distributional token overlap (top). The difference is especially stark for language pairs with different scripts (•), compared to language pairs with the same script (×). The orange line shows the linear fit across all included pairs.

1 Introduction

Highly multilingual language models have received plenty of research attention in recent years. *Cross-lingual alignment* of representations, that is, the similar representation of similar meanings regardless of input language (Libovický et al., 2020; Hämmerl et al., 2024), as well as good downstream cross-lingual transfer ability (cf. Huang et al., 2019; Schuster et al., 2019; Hu et al., 2020; Pham et al., 2024, etc.), have been considered desirable properties for such models. Representation alignment is typically seen as a key contributing factor to transfer ability, which in turn enables efficient handling of numerous task-language combinations. A number of papers have asked when and why information is shared across language boundaries in multilingual models and enables cross-lingual transfer (Dufter and Schütze, 2020; Deshpande et al., 2022; Limisiewicz et al., 2023; Hua et al., 2024; Schäfer et al., 2024, inter alia).

[1]https://github.com/KathyHaem/token-alignability

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 756–767
April 29 - May 4, 2025

Token overlap, i.e., the occurrence of identical tokens in the corpora of multiple languages, has been shown to affect the cross-lingual capabilities of models (Wu and Dredze, 2019). Another approach is to compare the distributions of token literals in parallel corpora (Limisiewicz et al., 2023). Still, both metrics have a crucial limitation: they cannot explain why related languages with different scripts are well-aligned by the models (see § 2.1).

Here, we propose another angle: token alignability. This concept captures the intuition that models may rely on statistical correspondences between subword tokens ('token alignment') that are more nuanced than literal string matching. From token alignments produced by a statistical word aligner, we derive two kinds of *token alignability scores* for any language pair in a multilingual tokeniser: one directional, one symmetrised (§ 3.2).

We compute correlations of these scores both to downstream transfer performance on classification and sequence labelling tasks (cf. § 3.3), and to measures of cross-lingual alignment in the model representations (cf. § 3.4). Our primary object of study is a set of small encoder models trained with several different multilingual tokenisers (BPE, Unigram, and 'TokMix'). Furthermore, we also consider recent larger, pre-trained decoder models. In addition to showing that token alignability is a better predictor of downstream cross-lingual transfer than distributional overlap (§ 4.1), we consider the impact of pre-training data size (§ 4.2), and show the correlation of token alignability with representation alignment inside the model (also § 4.1). Finally, we discuss how this insight may be applied to future multilingual tokenisers (§ 5).

2 Related Work

Subword tokenisation is currently the standard input processing approach of language models, with BPE (Sennrich et al., 2016) and UnigramLM (Kudo, 2018) being the most common algorithms for deriving these tokens. However, there has been increased interest in recent years in addressing limitations of the subword token paradigm (e.g., Alkaoud and Syed, 2020; Hofmann et al., 2022; Schmidt et al., 2024) or even moving beyond it (e.g., Xue et al., 2022; Mofijul Islam et al., 2022).

2.1 Influence of tokenisers on cross-linguality

Most relevant for our purposes are measurements of tokeniser properties (e.g., Zouhar et al., 2023; Batsuren et al., 2024), particularly for multilingual language models. Limisiewicz et al. (2023) measure the distance of a language pair's token vocabulary via divergence of the two token distributions. They find that this kind of 'soft overlap' measure correlates well with downstream transfer performance, with an important caveat: the observed correlations are strong for language pairs with the same script, but weaker for pairs with different scripts. This is because of how the metric is calculated: The occurrences of subword tokens are counted on each side of a parallel corpus, giving a distribution per language. Then, Jensen-Shannon-Divergence (**JSD**; Lin, 2006) is calculated, which gives a symmetrized distance between the two distributions of subword tokens. The literal matching limits the predictive power of their metric for pairs with different scripts—for instance, Hindi and Urdu are known to be related languages written in different scripts. Transfer between them works well, while the computed distance is large.

2.2 Word Alignment in MT

Alignment, in the sense used in statistical Machine Translation (MT) (Brown et al., 1993) is a mapping between parallel sentences, showing which tokens are translations of one another and how often they correspond across whole corpora. The original intuition behind attention is that it finds this kind of mapping in a contextualised manner (Bahdanau et al., 2015), whereas statistical word aligners (we use eflomal; Östling and Tiedemann, 2016) give a discrete mapping.

3 Methodology

Our central analysis relies on rank correlations, showing which tokeniser metrics (§ 3.1, § 3.2) are more predictive of downstream cross-lingual transfer (§ 3.3) and cross-lingual alignment of representations (§ 3.4). We ensure that within each task, the metrics are always compared over the same set of language pairs.

3.1 Distributional/Soft Overlap (JSD)

We measure soft overlap between the token distributions of two tokenised corpora. We follow the setting used by Limisiewicz et al. (2023) and outlined in § 2.1, but we compute it on the FLORES-200 corpus (Guzmán et al., 2019; Goyal et al., 2022; Team et al., 2022) for comparison with our proposed metrics. This score is symmetric between

both directions of a language pair. A lower score corresponds to a smaller distance and is thus better.

3.2 Token alignability of a language pair

We define the *token alignability score* for a language pair based on the symmetrised word alignment of one parallel corpus after training the tool on another. To train the priors, we use OPUS-100 data (Tiedemann, 2012; Zhang et al., 2020) for en-xx language pairs, and subsets of MultiCCAligned (Tiedemann, 2012; El-Kishky et al., 2020) for non-English language pairs. Seee Appendix A for a breakdown of language pairs. For each training corpus, we take up to 300k sentence pairs.

As our test corpus, we use FLORES-200 (Guzmán et al., 2019; Goyal et al., 2022; Team et al., 2022) because of its multi-parallel nature and less noise compared to MultiCCAligned. Following Vázquez et al. (2019), we run a statistical (discrete) word aligner (specifically **eflomal**; Östling and Tiedemann, 2016) on the test corpus with a single iteration. Based on the final symmetrised alignment over the test corpus, we can determine:

a) The *proportion of 1-1 token alignments* (higher is better), i.e., the rate of subword tokens in the source language text with a one-to-one correspondence to subword tokens of the target language text. We take this measure per direction, since it can be markedly lower if the source language is over-segmented.

b) The *eflomal score* (lower is better), which represents the tool's estimation of the "maximum unnormalized log-probability of links in the last sampling iteration" (Vázquez et al., 2019), given the learned priors over the subword vocabulary and corpus. We average this score over both directions of a language pair.

3.3 Downstream cross-lingual transfer

We were able to obtain model instances with several distinct tokenisers (BPE, Unigram, TokMix), and results for downstream cross-lingual transfer, from the authors of Limisiewicz et al. (2023). See Appendix B for brief model descriptions. This allowed us to run correlation analyses without retraining the models, instead testing our metrics against an existing set of experiments. The downstream results were obtained by fine-tuning the models on a given source language (any of the available languages for the task) and evaluating on a target language, resulting in many data points. The tasks tested are XNLI (Conneau et al., 2018), part-of-speech tagging (POS) and dependency tagging (UD) (both based on Zeman et al., 2019), and named entity recognition (NER; Pan et al., 2017). We always use Spearman's rank correlation to estimate the metrics' predictive power, following the previous work.

3.4 Cross-lingual embedding alignment

We measure cross-lingual alignment between a language pair as retrieval accuracy on the Tatoeba dataset (Artetxe and Schwenk, 2019) as well as the FLORES-200 development set. Following Jones et al. (2021), we additionally compute average margin distances on the latter, that is, how much closer the correct match is to the source sentence than other target-side sentences are. We do not compute word-level embedding alignment scores.

For encoder models, we create sentence embeddings by feeding the sentence to the model and averaging the encoder representations from layer 7 (with attention mask applied). The reasoning is that the middle layers in XLM-R and similar encoder models, such as the ones we use, have been found to be more cross-lingually aligned than the output layers (e.g. Muller et al., 2021). For decoder models, we follow Jiang et al. (2023) in using the prompt "This sentence: {sentence} means in one word:", then taking the last token representation of the last hidden layer as the sentence embedding.

4 Results and Discussion

4.1 Main results

Table 1 shows that eflomal score is better than JSD at predicting downstream transfer performance in the multilingual encoder models from Limisiewicz et al. (2023). This holds across all three tokenisation types, particularly for the word-level tasks. XNLI seems to behave differently, possibly because it is a sentence-level task in contrast with the other three, or because it has results available for fewer, mostly higher-resource, language pairs. Note also that XNLI transfer results were quite low in absolute terms.

Intuitively, JSD clusters language pairs with different scripts very closely together, even when they have markedly different transfer performance (see visualisations in App. Fig. 2–4). Eflomal score is not confounded by the different scripts, yielding better rankings within that group, and usually a better overall ranking. Meanwhile, the proportion of

Task	JSD			one-to-one			eflomal		
	all	=	≠	all	=	≠	all	=	≠
XNLI	-.33	-.57	**-.40**	.29	.50	.21	**-.45**	**-.60**	-.38
POS	-.45	**-.64**	-.45	.32	.36	.29	**-.64**	-.50	**-.64**
UD	-.23	-.25	-.25	.16	.33	.13	**-.41**	**-.36**	**-.42**
NER	**-.63**	-.25	**-.49**	.29	**.35**	.25	-.52	-.21	-.48

(a) Unigram

Task	JSD			one-to-one			eflomal		
	all	=	≠	all	=	≠	all	=	≠
XNLI	**-.55**	-.45	**-.40**	.11	**.46**	.05	-.44	-.39	-.29
POS	-.17	**-.65**	-.08	.35	.44	.33	**-.49**	-.52	**-.46**
UD	-.16	-.30	-.15	.18	.29	.19	**-.33**	**-.36**	**-.32**
NER	-.51	-.38	-.30	.30	**.53**	.28	**-.57**	-.25	**-.52**

(b) BPE

Task	JSD			one-to-one			eflomal		
	all	=	≠	all	=	≠	all	=	≠
XNLI	**-.45**	**-.44**	**-.43**	-.07	.34	-.23	-.36	-.43	-.22
POS	-.21	**-.69**	-.11	.11	.23	.06	**-.54**	-.51	**-.51**
UD	-.18	-.17	-.16	.01	.04	-.00	**-.38**	**-.33**	**-.39**
NER	-.38	**-.32**	-.09	.11	.23	.08	**-.48**	-.27	**-.42**

(c) TokMix

Table 1: Spearman's rank correlation of downstream transfer with JSD, proportion of one-to-one alignment, and eflomal score, for language pairs with the same (=) and with a different script (≠).

one-to-one alignments shows weaker or no correlation. This implies that the proportion of one-to-one alignments may be too simplistic here, while the eflomal score, as an estimate of log-probability, captures more nuance.

Table 2 lists correlations of JSD and eflomal score with three measures of embedding similarity (retrieval on Tatoeba and FLORES-200, and average margin on FLORES-200). These results are for the BPE model. The underlying distributions are shown in Fig. 5. We see that JSD gives clear correlations for all three measures in *same-script* language pairs, while eflomal score correlates more strongly on *different-script* language pairs.

All the correlations are much stronger on the FLORES dataset, likely because this dataset was used to calculate the tokeniser metrics in the first place. We can therefore see these as a kind of upper bound on how well the tokeniser metrics can predict cross-lingual alignment. The fact that the eflomal score is less predictive in the same-script group may indicate that the model does rely on more literal token matching when that information is available. To the extent that the behaviour differs from what is seen in Table 1, this underscores that cross-lingual embedding alignment, as measured by similarity, is just one factor in the cross-lingual transfer ability of the model.

Task	JSD			eflomal		
	all	=	≠	all	=	≠
F1 Flores	-.79	**-.70**	-.67	**-.83**	-.62	**-.81**
Avg mgn Flores	-.74	**-.72**	-.59	**-.80**	-.45	**-.79**
Tatoeba	**-.33**	**-.46**	-.19	-.33	-.27	**-.24**

Table 2: Spearman's rank correlation of embedding alignment with JSD and eflomal scores, on the BPE tokenizer/model. We show overall correlations (all), same-script (=), and different script (≠) pairs.

Model	XNLI	POS	UD	NER
Unigram	.87	.37	.33	.34
BPE	.80	.37	.49	.33
TokMix	.81	.34	.54	.26

Table 3: Rank correlation of downstream transfer from English with training size of the target language.

4.2 Is data size a confounder?

Table 3 shows data size in the trained encoders (and tokenizers), correlated with downstream transfer performance from English. Here, we consider only the pairs where English is the source language because English is generally the most dominant language, and there is some research suggesting that models "work" in English (Wendler et al., 2024). This correlates very well for XNLI, but much less in the other tasks. Again, XNLI stands out as a sentence-level task with fewer overall language pairs and relatively low transfer performance, so this result should be taken with a grain of salt. Overall, the correlations suggest that there is indeed a connection between data size and transfer ability, but data size cannot account for the whole effect. See also Table 6 in the Appendix.

4.3 What about decoders?

We additionally experiment with Mistral-7B-v0.1, Aya23-8B, and Llama-3-8B-Instruct, varying the model type, as well as the amount of multilinguality in pre- and post-training. For these, we calculate alignability scores, JSD, and representation alignment for a subset of language pairs. Table 4 shows rank correlation results. In Mistral, eflomal is still more predictive of overall representation alignment than JSD, while in Aya23 and Llama3, the opposite is true. This may suggest that cross-linguality in these decoder models works differently than in encoder models, or that they *do*

Model	Task	JSD			eflomal		
		all	=	≠	all	=	≠
Aya23	F1	**-.68**	**.31**	**-.73**	-.49	-.26	-.43
	Avg mgn	**-.65**	**.31**	**-.67**	-.43	-.26	-.36
LLaMA3	F1	**-.59**	-.26	**-.45**	-.32	**-.50**	-.18
	Avg mgn	**-.33**	-.74	**-.02**	-.21	**-.88**	-.02
Mistral	F1	-.20	-.05	.16	**-.59**	**-.67**	**-.55**
	Avg mgn	-.22	**.24**	.13	**-.74**	**-.24**	**-.76**

Table 4: Spearman's rank correlation of embedding alignment with JSD and eflomal scores, on decoders. We show overall correlations (all), same-script (=), and different script (≠) pairs.

rely more on literal token matches for their crosslinguality. Nevertheless, in Llama3-8B-Instruct, the eflomal score shows an unusually high correlation for same-script language pairs. Note also that absolute retrieval performance from the Mistral and Llama3 representations is quite low—Aya23 performs better. The corresponding visualisations are shown in Appendix C.4.

5 Future Work

We showed here that good tokeniser alignability correlates well with crosslinguality, an important factor for the performance of multilingual language models. Hence, the eflomal score may be applied to improve vocabulary learning for fairer multilingual tokenisers (see also Ahia et al., 2024; Limisiewicz et al., 2024). However, a naive implementation, where alignability score is checked at every decision point (merges for BPE, or pruning tokens for Unigram), is far too intensive. Therefore, future work in this area will require finding suitable approximations, like calculating alignability score difference for some fraction (e.g., on the order of 10%) of all candidate tokens at a time.

6 Conclusion

We have proposed a new metric for describing the quality of a multilingual tokenisation, with implications for cross-lingual alignment in multilingual pre-trained models: token alignability. This metric is particularly relevant for language pairs with different scripts and thus no literal token overlap. We showed correlations with transfer performance on downstream classification tasks, as well as with measures of cross-lingual alignment. These findings show the potential of our token alignability metric to guide the development of robust multilingual tokenisers and to identify suitable language pairs for cross-lingual transfer.

Limitations

Our study has focused on a relatively small set of models. We do not have extensive cross-lingual transfer experiments for decoder models because fine-tuning each model on any number of languages would take too much compute. Some of the downstream results from the previous work (particularly for XNLI) were quite poor in absolute terms, so they may not entirely reflect the situation in a higher-performance model. While alignability score for one language pair is not very time-consuming to compute (and can be done on CPU), the time adds up quickly for a broader set of language pairs. In its present formulation, alignability is also a corpus-wide score, meaning it would require reformulating for word-level tasks.

Acknowledgments

Thank you to Jindra Helcl for helpful discussions about this research. KH is supported by the Munich Center for Machine Learning, and did much of the work on this project during a research visit to Prague. The work at CUNI was supported by the Charles University project PRIMUS/23/SCI/023.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hoffman, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A. Smith. 2024. Magnet: Improving the multilingual fairness of language models with adaptive gradient-based tokenization. *preprint*, arXiv:2407.08818 [cs.CL].

Mohamed Alkaoud and Mairaj Syed. 2020. On the importance of tokenization in Arabic embedding models. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 119–129, Barcelona, Spain (Online). Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. 2024. Evaluating subword tokenization: Alien subword composition and oov generalization challenge. *preprint*, arXiv:2404.13292 [cs.CL].

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10922–10943, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *preprint*, arXiv:2003.11080 [cs.CL].

Tianze Hua, Tian Yun, and Ellie Pavlick. 2024. mOthello: When do cross-lingual representation alignment and cross-lingual transfer emerge in multilingual models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1585–1598, Mexico City, Mexico. Association for Computational Linguistics.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *preprint*, arXiv:2307.16645 [cs.CL].

Alexander Jones, William Yang Wang, and Kyle Mahowald. 2021. A massively multilingual analysis of

cross-linguality in shared embedding space. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5833–5847, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Taku Kudo. 2018. 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)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076, Bangkok, Thailand. Association for Computational Linguistics.

J. Lin. 2006. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, 37(1):145–151.

Md Mofijul Islam, Gustavo Aguilar, Pragaash Ponnusamy, Clint Solomon Mathialagan, Chengyuan Ma, and Chenlei Guo. 2022. A vocabulary-free multilingual neural tokenizer for end-to-end task learning. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 91–99, Dublin, Ireland. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106:125 – 146.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3184, Bangkok, Thailand. Association for Computational Linguistics.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. *preprint*, arXiv:2402.18376 [cs.CL].

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Anton Schäfer, Shauli Ravfogel, Thomas Hofmann, Tiago Pimentel, and Imanol Schlag. 2024. The role of language imbalance in cross-lingual generalisation: Insights from cloned language experiments. *preprint*, arXiv:2404.07982 [cs.CL].

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.

NLLB Team, 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, Loic 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. *preprint*, arXiv:2207.04672 [cs.CL].

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Raúl Vázquez, Umut Sulubacak, and Jörg Tiedemann. 2019. The University of Helsinki submission to the WMT19 parallel corpus filtering task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 294–300, Florence, Italy. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Daniel Zeman, Joakim Nivre, et al. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A Languages Included

We start from a set of 20 languages, namely the ones used by Limisiewicz et al. (2023) for their tokenizers: Arabic (ar), Turkish (tr), Chinese (zh), Greek (el), Spanish (es), English (en), Swahili (sw), Hindi (hi), Marathi (mr), Urdu (ur), Tamil (ta), Telugu (te), Thai (th), Russian (ru), Bulgarian (bg), Hebrew (he), Georgian (ka), Vietnamese (vi), French (fr), and German (de).

This gives us up to 190 language pairs (before accounting for direction), but we typically do not calculate numbers for *all* pairs, and each downstream task only has data available for some subset of the languages. We do compute all language pairs with English as either the source or target language. For non-English pairs, we compute token alignability for the product of these languages: ar, tr, zh, hi, ur, mr, ru, bg, vi, fr, es, ta, he.

B Encoder Details

The encoders were trained by Limisiewicz et al. (2023). The models' architecture is based on XLM-RoBERTa (Conneau et al., 2020). The size of the embeddings is 768, the number of attention layers is 8, and the number of attention heads is 6. The maximum sentence length is 128, and the vocabulary size in each tokenizer is 120000. The number of parameters is 150M, roughly half the size of XLM-R$_{base}$. See Limisiewicz et al. (2023) for training details. Their training corpus was a 10% subset of CC-100, with a balancing factor of $\alpha = 0.25$ (cf. Conneau and Lample, 2019). The model names BPE, Unigram, and TokMix are shorthand for their different vocabulary creation approaches. For BPE and Unigram, they simply applied the respective algorithm to the training set of all 20 languages, until reaching the target vocabulary size of 120000. For TokMix, they trained Unigram LM tokenisers for each language separately, and merged them by averaging token probabilities across tokenisers, then sorting and trimming. Our own experiments with these models were able to run on CPU.

C Additional Detail on Results

C.1 Graphs for Main Results

Figures 2, 3, and 4 visualise the distributions underlying Table 1. The sets of same- and different-script language pairs are colour-coded, and the overall correlations along with p-values are placed in the bottom left corner of each graph. Similarly, Figure 5 shows the distributions behind Table 2.

C.2 Analysis by Language Family

Similarly to our analysis of scripts, we assign language *pairs* to groups of same vs. different macro language families. We do this because some language families have just one representative in our set, while Indo-European accounts for many of the languages. We do not subdivide the macro language families for this analysis.

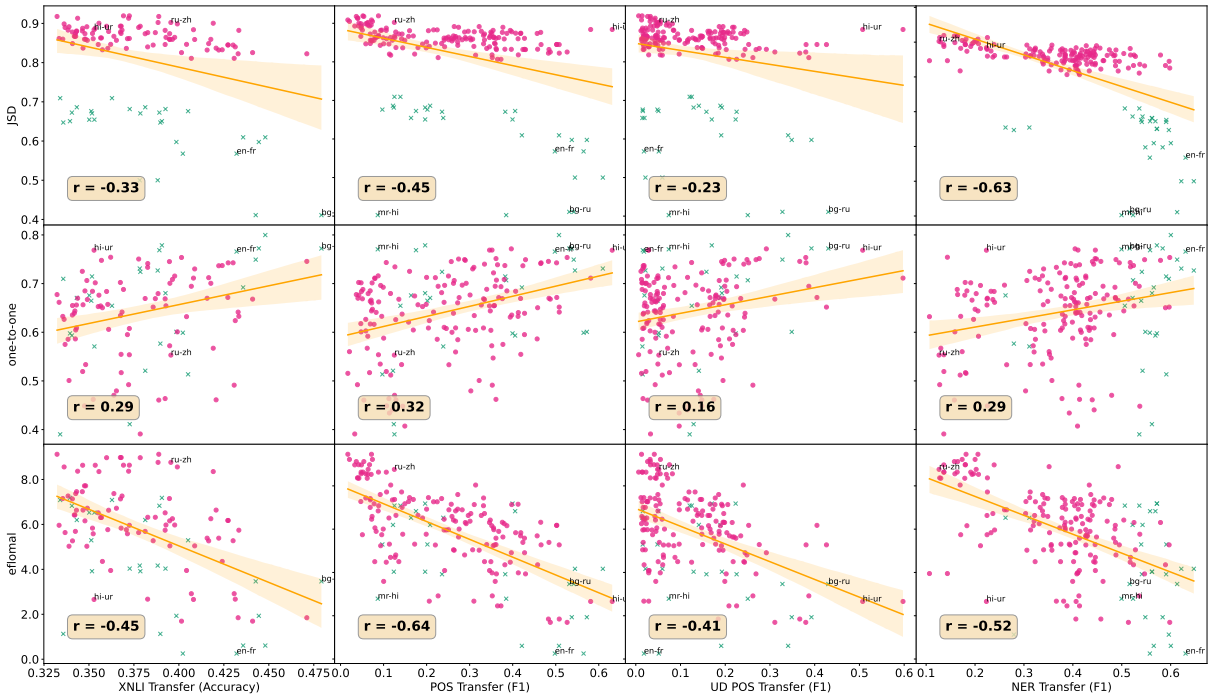

Figure 2: Unigram model: The eflomal score generally correlates better with downstream transfer than JSD. NER is the exception. Proportion of 1-1 token alignments, while it also breaks up the cluster of different-script language pairs, shows weaker or no correlations.

Task	Unigram			BPE			TokMix		
	all	=	≠	all	=	≠	all	=	≠
XNLI	-.38	-.60	-.22	-.29	-.34	-.26	-.22	-.42	-.23
POS	-.64	-.42	-.69	-.46	-.23	-.48	-.51	-.38	-.44
UD	-.42	-.30	-.41	-.32	-.08	-.37	-.39	-.33	-.33
NER	-.48	-.32	-.52	-.52	-.51	-.51	-.42	-.33	-.38

Table 5: Spearman's rank correlation of downstream transfer with JSD, proportion of one-to-one alignment, and eflomal score. This analysis shows only language pairs that use *different scripts*, further differentiated by whether they are in the same (=) or a different (≠) *language family*.

Table 5 shows the correlations of eflomal score with downstream cross-lingual transfer, over different-script pairs. We then split by same and different language families. In several cases, we see very similar correlations as on different-script pairs in general. XNLI stands out again, with pairs from the same language family tending to be more correlated across all tokenisers.

C.3 Data Size Correlated with Metrics

Table 6 shows the correlations of target language pre-training data sizes with our tokeniser metrics.

	JSD	one-to-one	eflomal
Unigram	-.30	.49	-.44
BPE	-.40	.24	-.54
TokMix	-.48	.30	-.52

Table 6: Spearman's rank correlation of the target language pre-training data size with our metrics. Only pairs with English as the source language are considered for this table.

C.4 Graphs for Decoder Results

The underlying distributions of Table 4 are visualised in Figure 6 for Aya23-8B, Figure 7 for Llama-3-8B-Instruct, and Figure 8 for Mistral. Both in Llama3-8B-Instruct and Aya23-8B, JSD correlates more strongly with cross-lingual alignment of representations, but all correlations here are weaker than is the case in the encoder models. For Mistral, eflomal score correlates more with cross-lingual alignment, which is in contrast to the other two decoder models.

Also, note that Aya23 shows decent retrieval performance, while the representations from Llama3 and Mistral both perform poorly on retrieval F1.

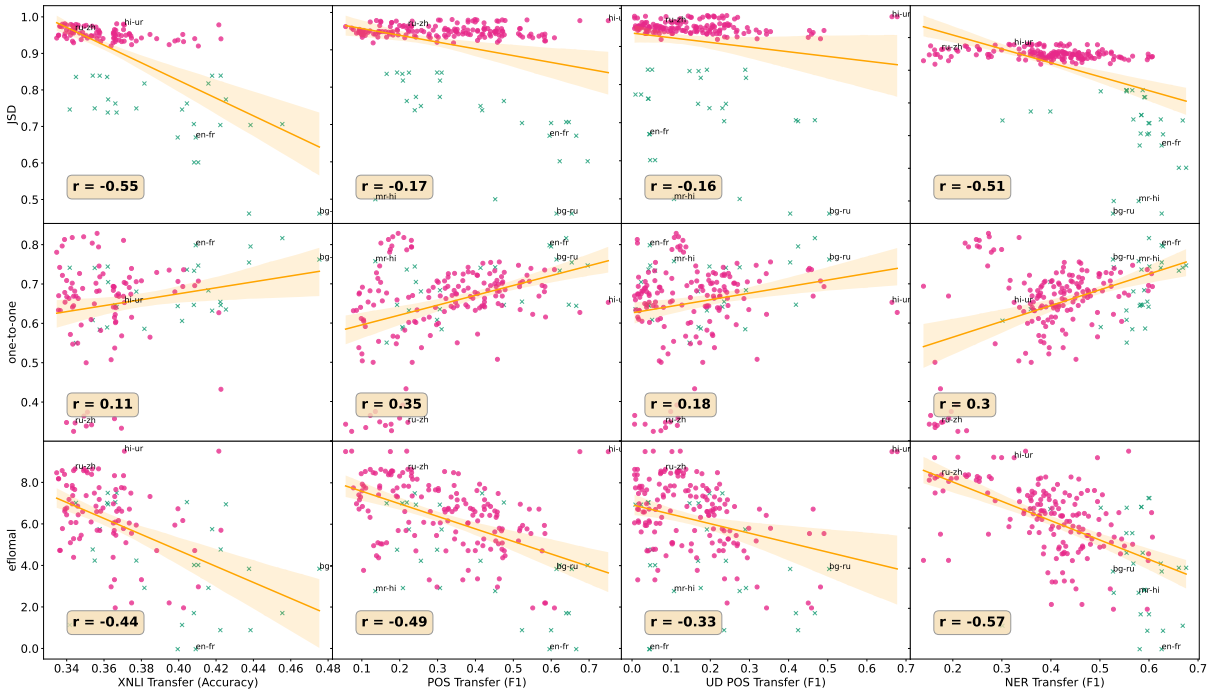

Figure 3: BPE model: The eflomal score correlates better with downstream transfer than JSD, with the exception of XNLI. Proportion of 1-1 token alignments, while it also breaks up the cluster of different-script language pairs, shows weaker or no correlations.

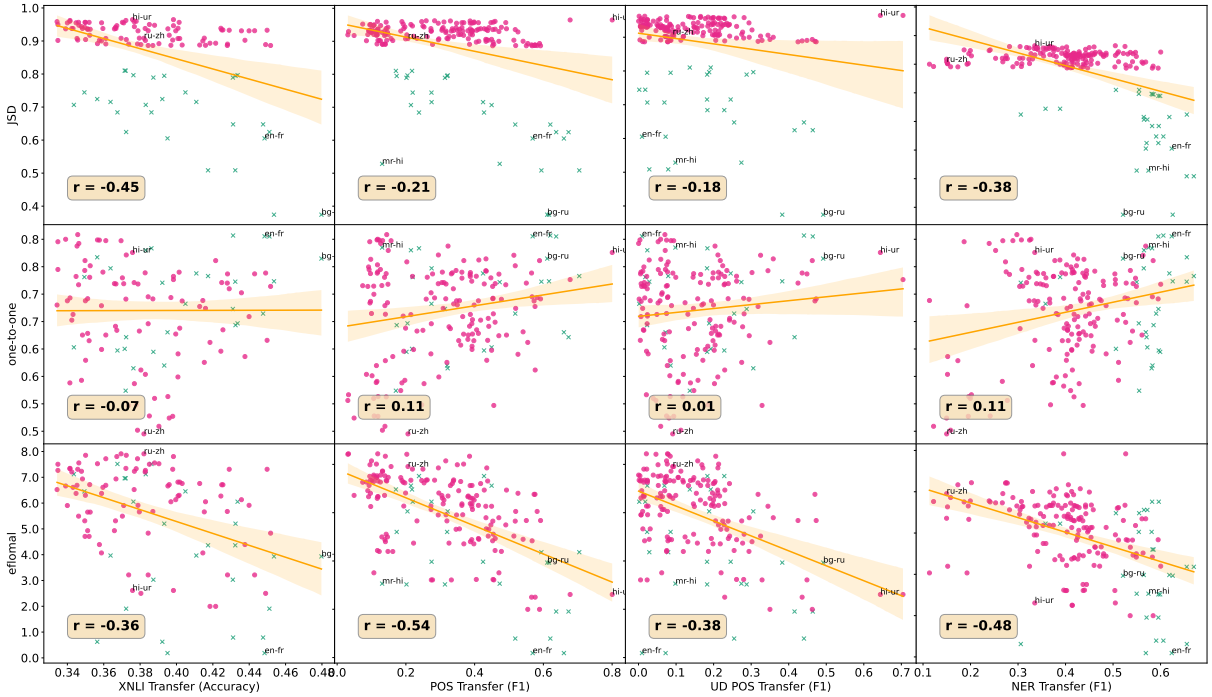

Figure 4: TokMix model: The eflomal score correlates better with downstream transfer than JSD, again with the exception of XNLI. Proportion of 1-1 token alignments, while it also breaks up the cluster of different-script language pairs, shows no correlations.

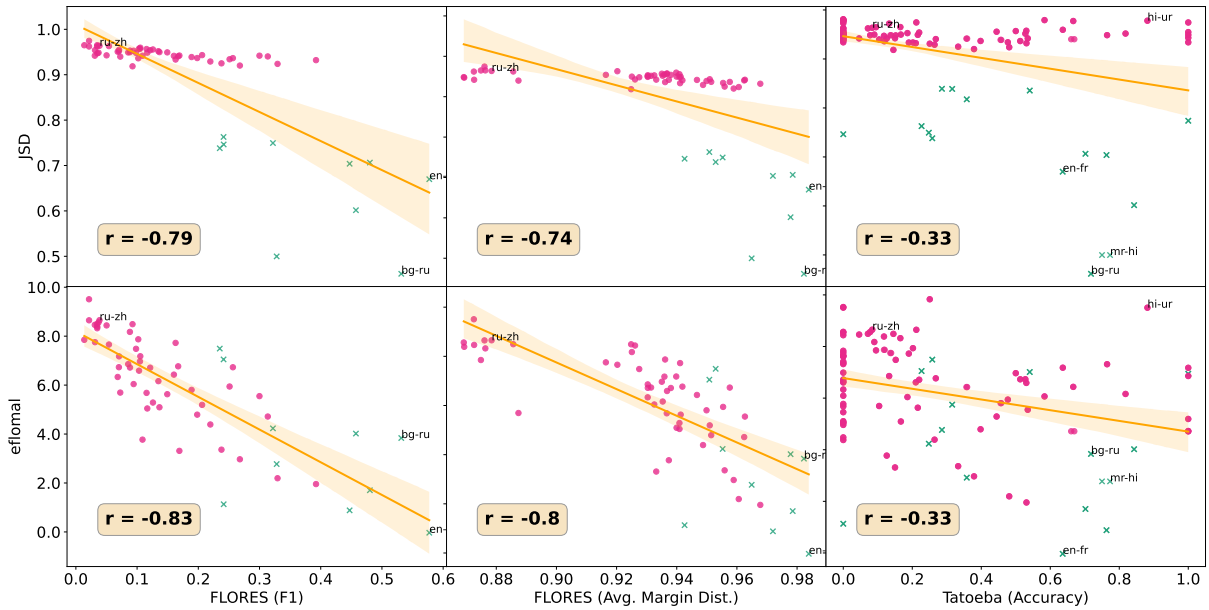

Figure 5: BPE Model: Eflomal scores correlates well with cross-lingual embedding alignment. Nevertheless, both metrics perform similarly over the Tatoeba dataset.

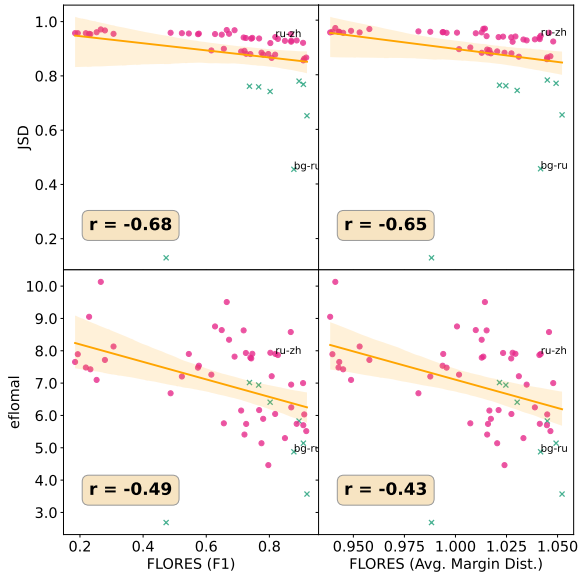

Figure 6: Aya23: Spearman's rank correlation of cross-lingual embedding alignment with JSD and eflomal score.

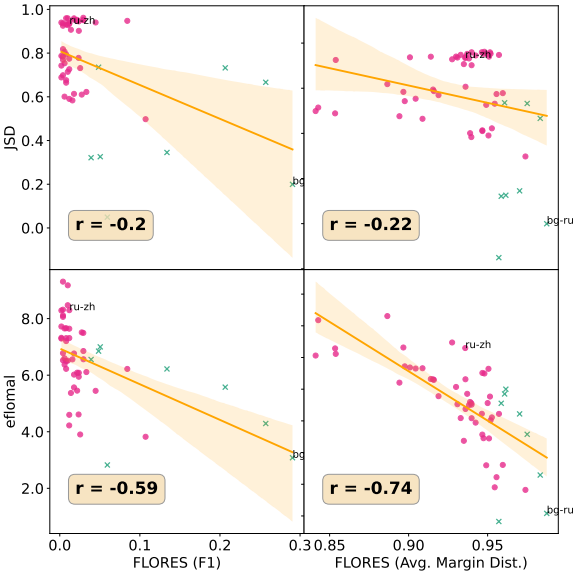

Figure 8: Mistral: Spearman's rank correlation of cross-lingual embedding alignment with JSD and eflomal score.

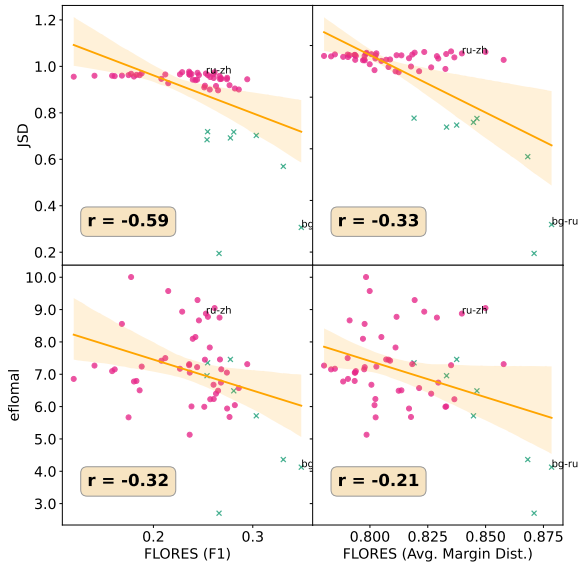

Figure 7: Llama3: Spearman's rank correlation of cross-lingual embedding alignment with JSD and eflomal score.