

Cross-Attention Eigenvalues Know the Word Order in Multilingual Machine Translation

Nikolai Prudnikov^{1,2}, Oleg Kachan¹

¹ HSE University, Moscow, Russia

² Sber AI Lab, Moscow, Russia

naprudnikov_1@edu.hse.ru, okachan@hse.ru

Abstract

Multilingual machine translation models must adapt to the grammatical structure of the target language, yet how structural information shapes decoder computation remains unclear. Using the eigenvalue spectra of per-(layer, head) attention Gram matrices instead of raw attention weights, we show that cross-attention geometry in NLLB-200 is organized around the target language: these spectral features carry a strong, distributed signal about target-language word order (SVO vs. SOV) when English is the source, with near-perfect decodability across cross-validation. The signal is not concentrated in a small subset of heads, but is broadly distributed across layers and heads—yet it degrades sharply at every level of aggregation when English is the target. This asymmetry suggests that cross-attention geometry reflects the demands of target-side generation rather than source-side structure. Leave-one-language-out evaluation confirms that the signal generalizes to unseen languages, ruling out language identity as a confound.

1 Introduction

Multilingual neural machine translation models must implicitly represent the structural properties of dozens or hundreds of languages simultaneously. Among the most fundamental of these properties is word order: whether a language places the verb before or after its object (SVO vs. SOV) has pervasive consequences for how source and target sequences align during generation. Locating this typological information within the model is essential for the interpretability of multilingual NLP models.

Beyond interpretability for its own sake, locating where target-language structure is encoded has practical stakes: it points to where cross-attention could be intervened on for language-specific fine-tuning, how to diagnose degraded generation quality for low-resource target languages, and whether a lightweight typology signal is recoverable from

attention geometry without decoding a full translation.

Prior work on probing transformer representations has focused predominantly on encoder-side models (Conneau et al., 2018; Belinkov, 2022) or on self-attention in encoder-decoder architectures (Raganato and Tiedemann, 2018). The cross-attention mechanism of the decoder, which mediates the interface between source representations and target generation, has received comparatively little attention as a locus of typological information. It is also unclear whether word-order signals, if present, are carried by specific layer-head pairs or distributed across the network.

To address these gaps, we probe the cross-attention mechanism of NLLB-200 (NLLB Team et al., 2022) across three model sizes and 22 typologically diverse languages. Our analysis relies on the eigenvalue spectra of attention Gram matrices, which capture the global shape of the attention distribution independently of specific token identities or alignments. This spectral representation is both stable and low-dimensional, making it a suitable target for probing without the confounds of raw attention weights. Our central hypothesis is that word-order information is encoded in the geometric structure of cross-attention, and that its distribution depends on the translation direction—specifically, whether the model generates into a varied target language or into fixed English. Our experiments confirm this hypothesis: the signal is strong and linearly decodable when English is the source, but severely degraded when English is the target, revealing that cross-attention geometry primarily reflects the demands of target-side generation rather than source-side structure.

Our implementation is available online.¹

¹<https://github.com/eiky0u/NLLB-Attention-Geometry-Probing>

2 Related Work

Probing transformer representations. The probing classifier framework (Conneau et al., 2018; Belinkov, 2022) has been widely used to study what linguistic information is encoded in neural model representations. Applied to NMT, Belinkov et al. (2017) showed that encoder representations capture morphological information, while decoder representations are weaker on this task. Raganato and Tiedemann (2018) probed encoder representations in transformer-based MT and found that higher layers encode more semantic information. Xu et al. (2021) showed that word translation occurs progressively in encoder layers, with cross-attention in the decoder serving primarily as a sequential alignment mechanism.

Attention analysis. Voita et al. (2019) showed that a small subset of self-attention heads in NMT encoders performs most of the useful work, with the remainder prunable without performance loss. Zhang et al. (2023) analyzed functional head specialization in a multilingual MT transformer and found that different language pairs share the same heads for the same functions, with deep cross-attention heads cooperating for word reordering. The interpretability of attention weights has been actively debated: Jain and Wallace (2019) argued that attention distributions are largely uncorrelated with feature importance, while Wiegrefe and Pinter (2019) contested this claim; Sun and Marasović (2021) further showed that raw weights conflate contributive and non-contributive components. Kim et al. (2021) found the most important heads for translation quality are shared across language pairs, suggesting limited functional specialization.

Word order in multilingual models. Nikolaev and Pado (2022) demonstrated that mBERT’s zero-shot performance on subordinate-clause detection is dominated by SVO/SOV typology. Otmakhova et al. (2022) showed that typologically different languages encode morphosyntactic features differently across BERT layers. Ji et al. (2023) showed that typologically similar languages share similar positional representations in a multilingual dependency parser, with SVO/SOV typology correlating with the structure of the positional space. Brinkmann et al. (2025) found that LLMs develop cross-lingual abstractions of morphosyntactic concepts across diverse languages.

Our work differs from all of the above in three key respects: we target the decoder’s cross-attention in a generative MT model rather than encoder representations, we rely on spectral geometry rather than raw attention weights or token-level alignments, and we explicitly control for translation direction to isolate source-side from target-side effects.

3 Experimental Setup

3.1 Model

We experimented with three sizes of NLLB-200 (NLLB Team et al., 2022): 600M, 1.3B, and 3.3B parameters. All models share the same encoder-decoder Transformer architecture and vocabulary, differing only in depth and size of hidden dimension. We focus on NLLB-200 as the largest publicly available massively multilingual encoder-decoder MT model, whose training is predominantly pivoted through English, making it well suited for probing the $\text{eng} \rightarrow X$ vs. $X \rightarrow \text{eng}$ asymmetry studied here.

3.2 Data

We used the FLORES-200 development set (NLLB Team et al., 2022), which provides professionally translated sentences across all 200 NLLB languages. We selected 22 languages paired with English: 11 SVO (French, German, Thai, Indonesian, Chinese, Russian, Vietnamese, Portuguese, Swahili, Finnish, Swedish) and 11 SOV (Japanese, Korean, Turkish, Persian, Hindi, Georgian, Tamil, Azerbaijani, Mongolian, Urdu, Bengali), giving 997 sentences per language. Word-order labels were taken from WALS (Dryer and Haspelmath, 2013).

We restrict the contrast to SVO vs. SOV because these are the only word-order categories attested at scale in NLLB-200 with stable WALS labels: VOS, OVS, and OSV are attested almost exclusively in endangered or extinct languages, none of which are represented in NLLB-200, while the handful of VSO languages present (e.g., Arabic, Irish) are too few to support a balanced comparison.

To isolate source- from target-side effects, we compare two translation modes: $\text{eng} \rightarrow X$ (English fixed as source, target varied across the 22 languages) and $X \rightarrow \text{eng}$ (English fixed as target, source varied).

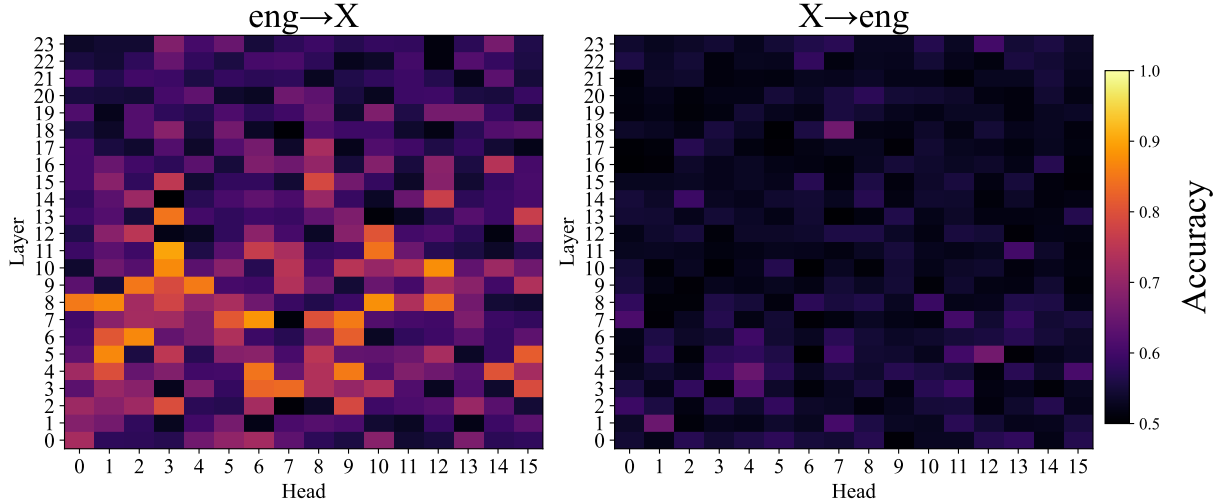


Figure 1: Probe accuracy per-(layer, head) averaged over 5 folds and 3 seeds for the 3.3B model. The $\text{eng} \rightarrow X$ map (left) shows a sparse but distinct signal, while the signal in the $X \rightarrow \text{eng}$ map (right) is virtually non-existent.

3.3 Feature Extraction and Probing

At each decoding step t , the cross-attention module produced a tensor of shape $[L, H, T, S]$, where L , H , T , and S denote the numbers of layers, heads, generated tokens, and source tokens. For each layer l and head h , we extracted the attention matrix $\mathbf{A} \in \mathbb{R}^{T \times S}$ and formed the Gram matrix $\mathbf{A}\mathbf{A}^\top$. Its eigenvalues are invariant to permutations of token order, allowing the spectrum to capture attention geometry independently of token identity or surface positional information. We computed the top- k eigenvalues, normalized them by their sum, and concatenated them across all $L \times H$ layer-head pairs to form the full feature vector, with subsets used for per-layer and per-head probes. We set $k = 12$, the smallest value retaining 98% of the cumulative normalized eigenvalue mass on average.

We trained a logistic regression probe (Belinkov, 2022) with standardized features (StandardScaler) to predict a binary label: $y = 1$ if the non-English language in the pair is SVO, $y = 0$ if SOV. A linear probe ensures that any recovered signal reflects structure in the features rather than classifier capacity.

We used two evaluation protocols: **cross-validation** with StratifiedGroupKFold (groups=sentence ID) with 5 folds averaged over 3 seeds, measuring generalization to unseen sentences from the same languages; and **leave-one-language-out** (LOO), where the probe is trained on 21 languages and evaluated on the held-out one, testing generalization to unseen languages and ruling out language identity as a confound.

Following Hewitt and Liang (2019), we additionally ran control tasks with randomly permuted word-order labels to verify that the probe is not merely memorizing labels and computed selectivity—defined as the probe’s accuracy on the true word-order task minus its accuracy on the control task with permuted labels.

4 Results

4.1 Cross-Validation

Cross-validation accuracy exceeded 0.999 across all three model sizes in the forward direction, and remained substantial when features were aggregated across all layer-head pairs in the reverse direction: 0.904 (600M), 0.906 (1.3B), and 0.919 (3.3B). Control-task selectivity is positive at the per-layer, per-head, and full-feature level in both translation directions, confirming that these accuracies reflect genuine word-order structure rather than label-fitting. Taken together, these results indicate that the word-order signal is robust across model scales and persists when the representations are considered at different levels of granularity.

Per-(layer, head) structure. Joint accuracy heatmaps showed no consistent structure: high-accuracy layer-head pairs were scattered across the grid without forming coherent horizontal or vertical bands (Figure 1). This suggests that word-order information is not concentrated in a dedicated subset of layer-head pairs, but is instead broadly distributed across the cross-attention mechanism, with different heads contributing to the signal at different layers.

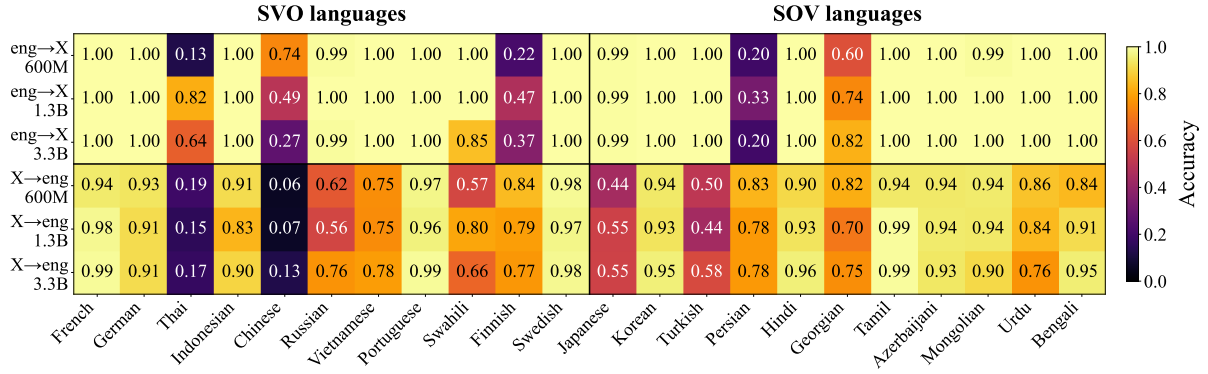


Figure 2: Leave-one-language-out accuracy for all six experimental conditions (three model sizes $\times$ two translation directions). Each cell shows the probe accuracy when the corresponding language is held out from training.

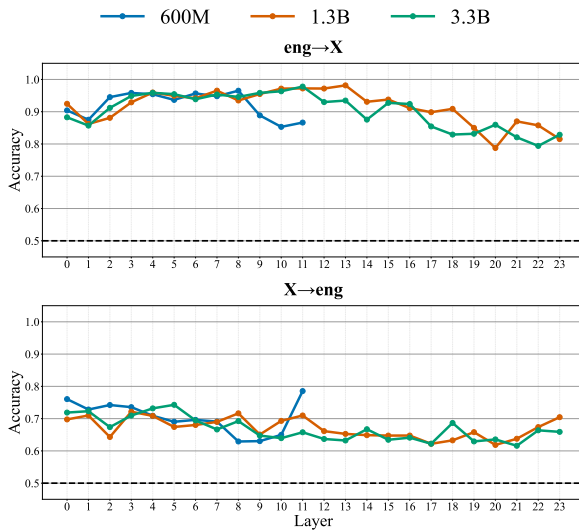


Figure 3: Layer-wise probe accuracy averaged over 5 folds and 3 seeds for eng→X (top) and X→eng (bottom) translation across three model sizes. Each point is a probe trained on the concatenated features of all heads within that layer, not an aggregation of per-head probes.

Per-layer profiles. In the forward direction, the per-layer accuracy profiles showed a clear depth-dependent structure (Figure 3). In the 600M model, accuracy remained high across all layers with a modest decline toward the final layers. In the 1.3B and 3.3B models, accuracy rose to a broad plateau in the middle layers, reaching a peak around layer 13, then degraded in the final layers, suggesting that word-order structure is most legible in the middle of the network. In the reverse direction, no pronounced mid-layer peak was observed: accuracy remained below 0.8 across all layers and declined gradually with depth. Overall, this suggests that the depth profile of the word-order signal depends on translation direction.

Per-head profiles. When features were aggregated from a single head across all layers, the forward-direction accuracy was uniformly high, with no head consistently lagging behind the mean. However, this uniformity at the per-head level does not imply specialization: the per-(layer, head) heatmap (Figure 1) shows that high-accuracy pairs are scattered across the network without forming coherent vertical or horizontal bands, indicating that the signal is distributed rather than concentrated in a dedicated subset. In the reverse direction, per-head accuracy was substantially lower, with most heads performing only modestly above chance and no head emerging as a consistent carrier.

4.2 Leave-One-Language-Out Analysis

The signal generalizes well to unseen languages: in the forward direction, most languages achieved accuracy above 0.99 across all model sizes (Figure 2). In the reverse direction, accuracy was lower but remained high for most languages, typically between 0.84 and 0.99. The outliers are linguistically interpretable and suggest that the probe tracks corpus-level word-order distributions rather than WALS typological labels.

Chinese and Thai were outliers in both translation directions. In the forward direction, Chinese accuracy degraded monotonically with model size, while Thai was unstable across models. In the reverse direction, both languages showed consistently low accuracy, consistent with their highly analytic, isolating structure, where individual tokens are less constrained to a single grammatical role. Japanese and Turkish achieved near-perfect accuracy in the forward direction but fell near chance in the reverse direction, consistent with the overall asymmetry.

Georgian improved monotonically with model size in the forward direction, suggesting that its complex morphosyntax requires greater model capacity to produce a geometrically distinctive signal.

In the forward direction, Finnish (SVO) was misclassified as SOV across all model sizes, while Persian (SOV) was misclassified as SVO. Both have structural properties that could plausibly disturb attention geometry independently of clause-level word order: Finnish’s rich agglutinative morphology packs information into single inflected forms, while Persian relies heavily on light-verb constructions, where a nominal element precedes a semantically bleached verb.

Finally, writing direction is not a confound. Persian (RTL) is a LOO outlier, but Urdu — also RTL and typologically similar — generalizes well. This confirms that the misclassification of Persian reflects its predicate structure rather than any artifact of right-to-left script processing.

5 Discussion

Asymmetry across translation directions. The signal is strongly asymmetric: robust and layer-structured in the forward direction ($\text{eng} \rightarrow \text{X}$), but degraded at every layer-head pair in the reverse ($\text{X} \rightarrow \text{eng}$). Since the same features and probe are used in both settings, and the effect persists under LOO evaluation, this asymmetry supports the hypothesis that cross-attention geometry is shaped by target-side generation. Producing a language with a particular word order induces structured attention geometry, whereas this influence is largely uniform when the target is English. This interpretation is supported by [Kudugunta et al. \(2019\)](#), who showed that encoder representations depend on the translation target. In $\text{eng} \rightarrow \text{X}$, the varying target induces variation in the (English) encoder representations that cross-attention exploits; in $\text{X} \rightarrow \text{eng}$, the target is fixed, removing target-driven variation across examples for a linear probe to detect.

What the eigenvalue spectrum encodes and why a linear probe works. The signal is recoverable from the spectrum of $\mathbf{A}\mathbf{A}^\top$ alone—without token identities or raw attention weights—indicating that word-order information is encoded in the global shape of the attention distribution rather than in specific alignments. The high accuracy of a linear classifier trained on these spectral features confirms that SVO and SOV targets yield systematically different spectral profiles. This is non-trivial: it im-

plies that the cross-attention dynamics produce a low-dimensional geometric summary that already separates typological classes linearly, without requiring non-linear transformations. Crucially, using the spectrum bypasses the known confounds of raw attention weights (which conflate multiple factors) and provides a more stable and interpretable signal.

Comparison with scalar attention statistics. Repeating LOO evaluation with four scalar features (entropy, confidence, coverage, variance) instead of eigenvalues reproduces the same $\text{eng} \rightarrow \text{X}$ vs. $\text{X} \rightarrow \text{eng}$ asymmetry, confirming it reflects a genuine property of attention distributions. However, the scalars differ in informativeness: entropy and confidence recover a stronger signal than coverage and variance, which are substantially less discriminative. Eigenvalues remain our primary choice, as they retain k -dimensional geometric characterization rather than collapsing the distribution to a single scalar.

Relation to prior work on attention heads. Our finding that the signal is distributed across many layer-head pairs, rather than concentrated in a few specialized heads, contrasts with observations on encoder self-attention ([Voita et al., 2019](#)), but agrees with findings that cross-attention heads in multilingual MT are not strongly language-pair specific ([Kim et al., 2021](#)). This suggests that cross-attention may operate as a collective mechanism, encoding structural information in the joint distribution across heads.

6 Conclusion

We investigated whether the geometry of cross-attention in NLLB-200 encodes target-language word order, using eigenvalues of per-(layer, head) Gram matrices as spectral features for a linear probe. We found that the signal is strong, distributed across the network, and driven by the target language: highly accurate when English is the source, it collapses at the layer-head level when English is the target. This asymmetry indicates that cross-attention geometry reflects target-side generation demands rather than source structure. Leave-one-language-out evaluation showed that generalization failures track corpus-level word-order distributions rather than typological labels.

7 Limitations

Our analysis is limited to a single model family, NLLB-200; whether this asymmetry generalizes to other multilingual MT architectures (e.g., M2M-100) or LLM-based translation systems remains an open question. We focus on SVO vs. SOV word order; whether cross-attention geometry also encodes other typological features (e.g., adposition type, adjective–noun order) is left for future work. The outlier status of analytic languages such as Chinese and Thai may partly reflect their shorter BPE-tokenized sequences, which could affect attention geometry independently of word order; we leave a systematic analysis of this confound to future work. Our evaluation relies solely on the FLORES-200 development set, which consists primarily of news-domain text; whether these findings generalize to other domains remains untested. Finally, our analysis is limited to 22 languages paired with English. Whether these findings generalize to non-English-pivoted translation pairs remains an open question for future work.

8 Ethics Statement

This work uses publicly available models (NLLB-200 licensed by CC-BY-NC 4.0) and datasets (FLORES-200 licensed by CC-BY-SA 4.0). No human subjects were involved. We foresee no ethical concerns.

9 Acknowledgements

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4E0002 and the agreement with HSE University No. 139-15-2025-009.

References

- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. [What do neural machine translation models learn about morphology?](#) In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.
- Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. [Large language models share representations of latent grammatical concepts across typologically diverse languages](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. [What you can cram into a single \$\\$ \& ! \# *\$ vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. [WALS Online](#). Max Planck Institute for Evolutionary Anthropology, Leipzig.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not Explanation](#). 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Tao Ji, Yuanbin Wu, and Xiaoling Wang. 2023. [Typology guided multilingual position representations: Case on dependency parsing](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13524–13541, Toronto, Canada. Association for Computational Linguistics.
- Zae Myung Kim, Laurent Besacier, Vassilina Nikoulina, and Didier Schwab. 2021. [Do multilingual neural machine translation models contain language pair specific attention heads?](#) In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2832–2841, Online. Association for Computational Linguistics.
- Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. [Investigating multilingual NMT representations at scale](#). 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.
- Dmitry Nikolaev and Sebastian Pado. 2022. [Word-order typology in multilingual BERT: A case study in subordinate-clause detection](#). In *Proceedings of*

the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 11–21, Seattle, Washington. Association for Computational Linguistics.

Eighth Conference on Machine Translation, pages 496–506, Singapore. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mehta Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.

Yulia Otmakhova, Karin Verspoor, and Jey Han Lau. 2022. [Cross-linguistic comparison of linguistic feature encoding in BERT models for typologically different languages](#). In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 27–35, Seattle, Washington. Association for Computational Linguistics.

Alessandro Raganato and Jörg Tiedemann. 2018. [An analysis of encoder representations in transformer-based machine translation](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Kaiser Sun and Ana Marasović. 2021. [Effective attention sheds light on interpretability](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4126–4135, Online. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Senrich, and Ivan Titov. 2019. [Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Sarah Wiegrefe and Yuval Pinter. 2019. [Attention is not not explanation](#). 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Hongfei Xu, Josef van Genabith, Qiuhui Liu, and Deyi Xiong. 2021. [Probing word translations in the transformer and trading decoder for encoder layers](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–85, Online. Association for Computational Linguistics.

Jingyi Zhang, Gerard de Melo, Hongfei Xu, and Kehai Chen. 2023. [A closer look at transformer attention for multilingual translation](#). In *Proceedings of the*