Interspeech 2024
1-5 September 2024, Kos, Greece

Wave to Interlingua: Analyzing Representations of Multilingual Speech Transformers for Spoken Language Translation

Badr M. Abdullah, Mohammed Maqsood Shaik, Dietrich Klakow

Language Science and Technology, Saarland University, Germany
Saarland Informatics Campus, Germany
{babdullah|mmshaik|dietrich}@lsv.uni-saarland.de

Abstract

In Transformer-based Speech-to-Text (S2T) translation, an encoder-decoder model is trained end-to-end to take as input an untranscribed acoustic signal in the source language and directly generate a text translation in the target language. S2T translation models can also be trained in multilingual setups where a single front-end speech encoder is shared across multiple languages. A lingering question, however, is whether the encoder represents spoken utterances in a language-neutral space. In this paper, we present an interpretability study of encoder representations in a multilingual speech translation Transformer via various probing tasks. Our main findings show that while encoder representations are not entirely language-neutral, there exists a semantic subspace that is shared across different languages. Furthermore, we discuss our findings and the implication of our study on cross-lingual learning for spoken language understanding tasks.
Index Terms: Transformer models, speech translation, interpretability, multilinguality

1. Introduction

Transformer-based architectures have witnessed a rapid rise in their adoption for various speech processing tasks due to their inherent flexibility and scalability. One of the main advantages of Transformer-based speech models is that they can efficiently be pretrained on a large corpus of untranscribed speech via self-supervised objectives, either monolingually [1, 2] or multilingually [3, 4]. Multilingual pretraining of speech models has been shown to facilitate cross-lingual transfer for low-resource languages [4, 5, 6, 7].

The multilingual capabilities of text-based, multilingual language models (MLMs) such as Multilingual BERT [8] and XLM-RoBERTa [9] have been empirically explored in numerous previous studies. One research direction seeks to unveil the zero-shot cross-lingual transferability of these models [10, 11, 12]. Another line of research focuses on the interpretability of their cross-lingually shared space, or the so-called 'interlingual" space [13, 14, 15, 16]. Earlier studies postulated that shared subword tokens are pivotal in aligning representations across languages [10, 17, 18]. However, considering that different languages exhibit structural similarities at various linguistic levels, there is increasing evidence indicating that MLMs uncover deep linguistic patterns in the data, leading to the cross-lingual transferability of their representations [19, 20, 21, 22].

Similarly, there has been a considerable interest in interpreting shared semantic spaces of text-based systems for multilingual machine translation (NMT) [23, 24, 25]. It has been observed that text encoder representations in NMT models usually form clusters that are based on linguistic similarity [26]. On the other hand, multilingual encoders of direct speech translation models remain unexplored with little known about the cross-lingual nature of their representations. We are aware of only one published study on interpreting the latent representations of multilingual S2T translation models [27]. Contrary to text-based encoders, which have access to acoustically-invariant written sentences, speech encoders have to deal with the inherent variability of spoken language. That is, speech encoders need to process the acoustic signal to extract the linguistic message, filtering out background noise and abstracting away from speaker and gender variability. The complex nature of the input in speech translation requires additional auditory processing tasks that are not required for text-based NMT systems, which adds to the analytical value of speech translation encoders from both a scientific and engineering point of view.

In this paper, we present the first study to *probe* the representation of **multilinguality** in multilingual speech translation models that are only trained on acoustic data as input without text-based grounding at the source side. Our main findings are the following:

- The spoken language can be *linearly* identified from encoder representations, which indicates that they are not truly "interlingual", even in the higher layers of the encoder (§4).
- Despite the fact that encoder representations are not entirely language-neutral, the encoder exhibits a *shared semantic subspace* that enables spoken translation retrieval across different languages (§5).

Furthermore, we discuss the implications of our research and future directions that could be built on our findings in this paper.

2. Preliminaries

2.1. Speech Translation: Problem Definition

In this section, we formally describe the speech translation task. Consider a spoken utterance that consists of T acoustic frames $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_T) \in \mathcal{X}$, where each acoustic frame is a short-time temporal window that is assumed to be a quasi-stationary signal. Typically, speech signals are segmented into consecutive windows at the rate of 50 frames per second. In end-to-end Transformer-based speech translation, the encoder builds up a contextualized representation $\mathbf{c} = (\mathbf{c}_1, \cdots, \mathbf{c}_T)$ for each frame via the attention mechanism as follows

$$\mathcal{E} : (\mathbf{x}_1, \cdots, \mathbf{x}_T) \quad \mapsto \quad \mathbf{c} \in \mathbb{R}^{D \times T} \tag{1}$$

where D is the dimensionality of the last layer encoder representation. Usually, the last encoder representations are transformed via an adapter function[1] to reduce the sequence length and fit the decoder dimensionality as follows

[1] a parametrized feed-forward neural network.

10.21437/Interspeech.2024-2109

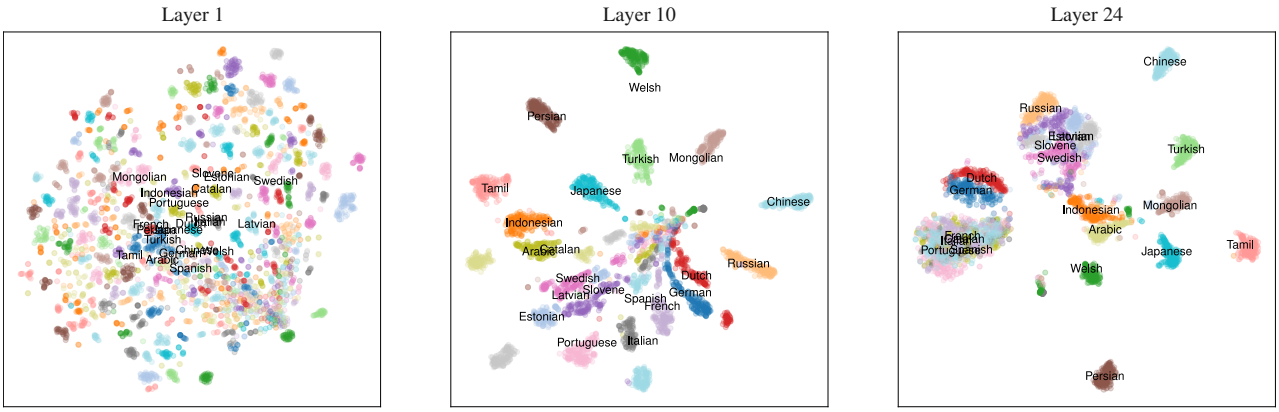

Figure 1: *T-SNE projection of utterance representations from encoder layers 1 (left), 10 (middle), and 24 (right).*

$$\mathcal{A} : (\mathbf{c}_1, \cdots, \mathbf{c}_T) \quad \mapsto \quad \hat{\mathbf{c}} \in \mathbb{R}^{d \times \tau} \tag{2}$$

where τ is the reduced sequence length such that $\tau < T$. The decoder then uses cross-attention to access the encoder contextualized representations $\hat{\mathbf{c}} = (\hat{\mathbf{c}}_1, \cdots, \hat{\mathbf{c}}_\tau)$ and generates a word sequence as follows

$$\mathcal{D} : (\hat{\mathbf{c}}_1, \cdots, \hat{\mathbf{c}}_\tau) \quad \mapsto \quad \mathbf{y} \in \mathcal{Y} \tag{3}$$

where $\mathcal{Y}$ is the word sequence space and $\mathbf{y} = (y_1, \cdots, y_N)$ is the text translation in the target language. Note that in this paradigm, the model does not produce any text transcription of the input in the source language.

2.2. Research Hypotheses

Given that spoken language is notoriously variable, speech translation encoders need to successfully perform auditory processing tasks that text-based encoders do not need to perform. These tasks include, for example, noise separation, speaker normalization, and spoken word detection. In order to understand the behavior of the multilingual model and the role of its components to perform the task, we formulate two different hypotheses: (1) The encoder of S2T translation models specializes in low-level auditory processing tasks, while the decoder takes care of high-level linguistic processing and utterance understanding, and (2) The encoder actively participates in understanding the meaning of an utterance, which yields a semantic space in its representations. We refer to hypothesis (1) as the **modular** hypothesis, where the encoder behaves as a speech recognition encoder that lacks deep semantic understanding of the input. On the other hand, we refer to hypothesis (2) as the **integrative** hypothesis, where the encoder and decoder share the task of utterance understanding, implying that the encoder exhibits a semantic subspace that is shared across languages.

3. Speech Translation Model

For this study, we use the multilingual S2T translation model developed by [4], which is based on wav2vec2-XLS-R-128 with 300M parameters, 24 encoder layers, and further fine-tuned on CoVoST-2 speech translation dataset [28]. Since we focus on the cross-lingually shared semantic space in the encoder representations, we examine the model that translates from 21 languages into English. The supported languages are: Arabic (ar), Catalan (ca), Welsh (cy), German (de), Estonian (et), Persian (fa), Indonesian (id), Japanese (ja), Latvian (lv), Mongolian (mn), Slovenian (sl), Swedish (sv), Tamil (ta), Turkish (tr), Chinese (zh), Spanish (es), French (fr), Italian (it), Dutch (nl), and Portuguese (pt). Based on how well they are represented in the CoVoST-2 dataset, these languages range from high-resource languages (e.g., French and German) to low-resource languages (e.g., Tamil and Japanese). We do not analyze the decoder representations of this model in this paper.

4. Experiment I: Language Identification

In this experiment, our goal is to better understand the representational geometry of multilingual space in the S2T translation model and how it evolves across the layers of the encoder. Note that the encoder in this multilingual model is not given any explicit signal that specifies the language of the input utterance. Therefore, we are interested in the question: **Do the encoder representations implicitly encode language identity?**

4.1. Experimental Data and Setup

For this experiment, we use a subset of in-domain data from the CoVoST-2 dataset. Each utterance is transformed into a single vector representation in each layer via mean pooling over frame representations.

4.2. Exploratory Visualization Analysis

First, we use the t-SNE dimensionality reduction algorithm to explore the emergent language space in the multilingual S2T translation encoder. Our observation is that utterances are not grouped by the respective language in the initial layers, but a language-specific cluster structure emerges in deeper layers. In Fig. 1, we depict the utterance representations in three crucial layers: layer 1, layer 10, and layer 24. Note that the name of the language is rendered at the centroid of its respective cluster. We observe a strong cluster structure in layer 10, where each language tends to have its own cluster. Interestingly, clusters of closely related languages seem to merge in last encoder layer (i.e., layer 24). For example, the Latin-based, Romance languages (i.e., Spanish, Catalan, Portuguese, Italian, and French) form a single coherent cluster in the last layer which was not observed in layer 10. Some of the other multilingual clusters cannot be explained by historical language relatedness but rather a geographic proximity and cultural similarity of the speaker communities as in the case of Latvian and Estonian, a Balto-Slavic

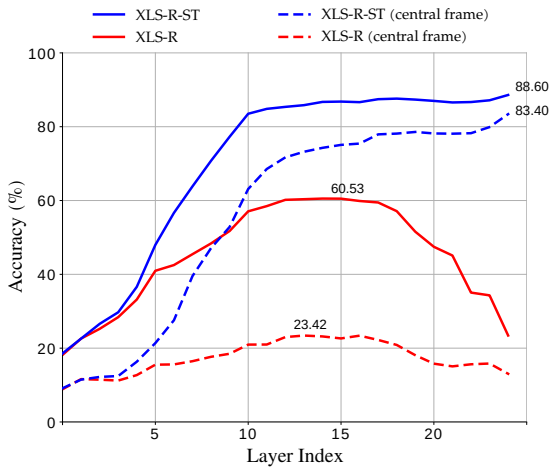

Figure 2: *Language identification via linear probing classifiers across the encoder layers. XLS-R-ST refers to the model that was fine-tuned on speech translation task.*

Indo-European language and Finno-Ugric Uralic language, respectively.

4.3. Probing via Linear Classifiers

Next, we use diagnostic classifiers to identify the language of the utterance from its representation. To do so, we train simple logistic regression-based classifiers that aim to predict the spoken language. For this experiment, we compare the classification performance of the translation model to that of the pre-trained XLS-R-128 model. We also investigate the effect of contextualization in predicting the language in both models by training classifiers on the central frame of the utterance.

The results of this probing experiment are shown in Fig. 2. First, we observe that the language is more linearly identifiable in the speech translation encoder compared to the pre-trained encoder but not in the pre-trained model, as shown by the blue vs. red dashed lines. Note that the inverted U-shape of the pre-trained encoder in Fig. 2 (the solid red line where performance peaks in middle layers) is a consequence of the so-called autoencoder-style behavior of the pre-trained wav2vec 2.0 architecture. That is, it has been observed in prior studies that middle layers are better at capturing linguistic units that require a higher level of contextualization such as phonemes and words [29]. Fine-tuning the pre-trained encoder on the speech translation task seems to strengthen the level of contextualization across the layers and changes the autoencoder-style behavior of the Transformer encoder. Thus, the language identification performance peaks at the last encoder layer. Also, we observe that training classifiers on the central frame of the utterance—instead of taking the mean across all frames—yields strong language identification performance in the S2T translation encoder, but not in the pre-trained model (blue vs. red dashed lines). These results suggest that deeper layers in speech translation models exhibit a higher level of contextualization that enables the identification of the spoken language even from a single frame of the utterance representation.

The results presented in §4.2 and §4.3 suggest that the geometry of the encoder representations in the S2T translation model is mainly organized by the language of the spoken input. Precisely, the representations in deeper layers tend to exhibit a language-specific cluster structure where the language can be identified using a linear classifier. Therefore, we conclude that the representation space does not "factor out" the language in a way such that the nearest neighbour to an utterance is the most semantically similar utterance in another language.

5. Experiment II: Shared Semantic Space

In the previous experiment, our goal was to analyze the degree to which language identity influences the geometry of the encoder representations. Since our findings suggest that the cluster structure of the representation space is mainly shaped by the language identity, one could conclude that the representations are not aligned according to their meaning. However, this does not exclude the existence of a **weakly aligned** representation space. That is, even though a semantic equivalent of an utterance may not always be the nearest neighbor of all data points in the representation space, it could still be the nearest neighbor if we filter out utterances by a specific, different language. In this experiment, we are interested in the question: **Is there a semantic subspace where the encoder representations are cross-lingually aligned?**

5.1. Experimental Data and Setup

In order to examine the semantic subspace that is shared across languages, we need a multilingual speech data that is aligned at the utterance level. To this end, we use the FLEURS speech dataset, which is an n-way parallel speech dataset in more than 100 languages, produced by native speakers of the respective languages [30]. Similar to Experiment I, each utterance is transformed into a single vector representation in each layer via mean pooling over frame representations.

5.2. Spoken Translation Retrieval

In this probing task, the goal is to take a query utterance in language A and retrieve its spoken translation in language B. Concretely, we compute the distance in the representational space between the query representation and all utterance representations in language B. Then, the utterance in language B with the smallest distance will be considered a candidate translation of the query. If the candidate is indeed the correct translation, then the model gets a score of $+1$, otherwise 0. Finally, once this procedure has been applied to all queries, we normalize by diving by the total number of query utterances to obtain a value $\in [0, 1]$. This corresponds to the precision@1 metric in information retrieval tasks. Note that this task is completely non-parametric. That is, we do not learn any additional weights on the top of the model representations to yield a better cross-lingual alignment.

We conduct the spoken evaluation task using the FLEURS speech dataset, focusing on language pairs from high-resource languages. The decision to focus on high-resource languages is driven by the fact there is a big gap in translation performance between high-resource and low-resource languages. Therefore, we do not expect any successful alignment between, for example, Tamil and Japanese, since the model is poor at translating utterances from these languages to English.

The result of this probing task is shown in Fig. 3. We focus on six language pairs: German $\rightarrow$ Dutch, Catalan $\rightarrow$ French, German $\rightarrow$ Italian, Dutch $\rightarrow$ German, German $\rightarrow$ French, and Russian $\rightarrow$ French. In the notation A $\rightarrow$ B, A denotes the language of the query utterance, while B denotes the language of the search utterances from which the translation is to be retrieved. To put our results in perspective, we benchmark against the

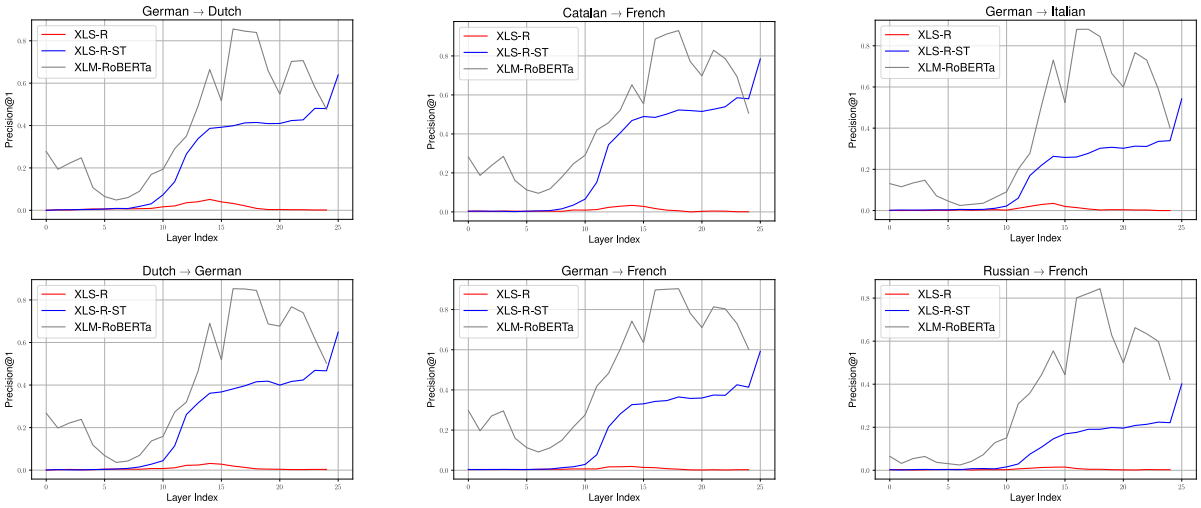

Figure 3: *Retrieval performance of the spoken translation retrieval task measured by the precision@1 metric.*

text-based, multilingual XLM-RoBERTa model, which serves as an upper bound for retrieval performance. This model is an encoder-only MLM, featuring the same number of layers (i.e., 24 layers) as the speech translation model analyzed in our study. The XLM-RoBERTa's evaluation is conducted using the text transcriptions of the spoken utterances. To establish a baseline and lower bound of the retrieval performance, we include the pre-trained self-supervised XLS-R-128 model in the evaluation. For this probing task, we also incorporate the adapter layer, denoted as layer 25 in the figure. In Fig. 3, we first observe that the pre-trained XLS-R-128 model is very poor on this task despite the multilingual nature of its pre-training corpus. This indicates that the pre-training masked language modeling objective in wav2vec 2.0 does not provide enough supervision signal for the model to learn a semantic space that is shared across languages. On the other hand, we observe an expected higher performance by XLM-RoBERTa model that peaks between layers 16-18 for all language pairs. What is interesting to observe here is the performance of the S2T translation encoder, which seems to be similar to the pre-trained speech model between layer 0-8. Yet, from layer 9 onwards, the speech translation encoder begins to diverge from the pre-trained model's trajectory. These observations align with previous research findings indicating that fine-tuning has minimal impact on the initial layers of Transformer-based speech models [29]. Moreover, the retrieval performance substantially improves in the adapter layer representation relative to the last encoder layer, suggesting that this layer specializes in interlingual, high-level linguistic processing. Another observation is that language similarity seems to be a strong effect of the spoken translation retrieval performance. That is, retrieval performance is higher for related language pairs (e.g., Catalan → French) compared to unrelated language pairs (e.g., Russian → French). This suggests that representations of related language pairs exhibit a stronger form of alignment.

In summary, the results of Experiment II demonstrate that the multilingual speech encoder indeed exhibits a semantic space that is cross-lingually shared. While the cross-lingual alignment may represent a weaker form of alignment, the strength of this alignment seems to be greatly influenced by the typological similarity between languages.

6. General Discussion

In this work, we aim to test two different hypotheses of information processing that describe the behavior of the inner workings of multilingual speech translation models and the contribution of the encoder to this task. The **modular** hypothesis posits that the encoder should behave like an ASR component, merely building an abstract speaker-invariant representation of the speech sounds in the input. On the other hand, the **integrative** hypothesis postulates that the encoder is not merely an ASR front-end, but rather an active component that engages in higher-level linguistic processing of the input.

The results of our experiments provide evidence in favor of the integrative hypothesis. A cross-lingual alignment between semantically related utterances of different languages is only possible if the encoder learns a semantic subspace where utterances that denote the same concept are projected nearby in space. Precisely, while the lower layers seem to perform auditory processing tasks that are necessary for understanding spoken content, deep layers in the multilingual speech encoder build representations that are of a semantic nature. The implication of our findings is that zero-shot and few-shot cross-lingual learning is highly feasible if we leverage pre-trained speech translation encoders for spoken language understanding tasks.

7. Conclusion

In this paper, we presented an interpretability study on the Transformer encoder representations for multilingual speech translation. Although our experiments show that encoder representations are not entirely language-neutral, the encoder exhibits a shared semantic subspace that enables spoken translation retrieval across different languages.

8. Acknowledgements

We thank the anonymous reviewers for their positive feedback. We sincerely thank Bernd Möbius for his valuable comments on the work presented in this paper. We extend our thanks to Gaofei Shen, who inspired us with the title of this paper. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

9. References

[1] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[3] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[4] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, H. Ko and J. H. L. Hansen, Eds., pp. 2278–2282.

[5] N. San, M. Bartelds, M. Browne, L. Clifford, F. Gibson, J. Mansfield, D. Nash, J. Simpson, M. Turpin, M. Vollmer, S. Wilmoth, and D. Jurafsky, "Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 13-17, 2021*. IEEE, 2021, pp. 1094–1101.

[6] S. Khurana, N. Dawalatabad, A. Laurent, L. Vicente, P. Gimeno, V. Mingote, and J. Glass, "Cross-lingual transfer learning for low-resource speech translation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[7] N. San, G. Paraskevopoulos, A. Arora, X. He, P. Kaur, O. Adams, and D. Jurafsky, "Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens," in *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, M. Hahn, A. Sorokin, R. Kumar, A. Shcherbakov, Y. Otmakhova, J. Yang, O. Serikov, P. Rani, E. M. Ponti, S. Muradoğlu, R. Gao, R. Cotterell, and E. Vylomova, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 100–112. [Online]. Available: https://aclanthology.org/2024.sigtyp-1.13

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of ACL*. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451.

[10] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proceedings of the ACL*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4996–5001. [Online]. Available: https://aclanthology.org/P19-1493

[11] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," in *Proceedings of EMNLP-IJCNLP*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 833–844.

[12] K. Karthikeyan, Z. Wang, S. Mayhew, and D. Roth, "Cross-lingual ability of multilingual bert: An empirical study," in *International Conference on Learning Representations*, 2019.

[13] E. A. Chi, J. Hewitt, and C. D. Manning, "Finding universal grammatical relations in multilingual BERT," in *Proceedings of ACL*. Online: Association for Computational Linguistics, Jul. 2020, pp. 5564–5577.

[14] J. Libovický, R. Rosa, and A. Fraser, "On the language neutrality of pre-trained multilingual representations," in *Findings of EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1663–1674. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.150

[15] A. Conneau, S. Wu, H. Li, L. Zettlemoyer, and V. Stoyanov, "Emerging cross-lingual structure in pretrained language models," in *Proceedings of ACL*. Online: Association for Computational Linguistics, Jul. 2020, pp. 6022–6034. [Online]. Available: https://aclanthology.org/2020.acl-main.536

[16] H. Gonen, S. Ravfogel, Y. Elazar, and Y. Goldberg, "It's not Greek to mBERT: Inducing word-level translations from multilingual BERT," in *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, Nov. 2020, pp. 45–56. [Online]. Available: https://www.aclweb.org/anthology/2020.blackboxnlp-1.5

[17] S. Cao, N. Kitaev, and D. Klein, "Multilingual alignment of contextual word representations," in *International Conference on Learning Representations*, 2019.

[18] J. Singh, B. McCann, R. Socher, and C. Xiong, "BERT is not an interlingua and the bias of tokenization," in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 47–55. [Online]. Available: https://aclanthology.org/D19-6106

[19] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," in *Proceedings of ACL*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4623–4637. [Online]. Available: https://aclanthology.org/2020.acl-main.421

[20] I. Papadimitriou and D. Jurafsky, "Learning Music Helps You Read: Using transfer to study linguistic structure in language models," in *Proceedings of EMNLP*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6829–6839.

[21] P. Dufter and H. Schütze, "Identifying elements essential for bert's multilinguality," in *Proceedings of EMNLP*, 2020, pp. 4423–4437.

[22] G. Shen, M. Watkins, A. Alishahi, A. Bisazza, and G. Chrupała, "Encoding of lexical tone in self-supervised models of spoken language," *arXiv preprint arXiv:2403.16865*, 2024.

[23] C. Espana-Bonet, A. C. Varga, A. Barrón-Cedeno, and J. Van Genabith, "An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1340–1350, 2017.

[24] H. Schwenk, D. Kiela, and M. Douze, "Analysis of joint multilingual sentence representations and semantic k-nearest neighbor graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6982–6990.

[25] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, 2019.

[26] S. Kudugunta, A. Bapna, I. Caswell, and O. Firat, "Investigating multilingual NMT representations at scale," in *Proceedings of EMNLP-IJCNLP*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1565–1575. [Online]. Available: https://aclanthology.org/D19-1167

[27] H. Sun, X. Zhao, Y. Lei, S. Zhu, and D. Xiong, "Towards a deep understanding of multilingual end-to-end speech translation," in *Findings of EMNLP*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14 332–14 348.

[28] C. Wang, A. Wu, J. Gu, and J. Pino, "Covost 2 and massively multilingual speech translation." in *Interspeech*, 2021, pp. 2247–2251.

[29] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[30] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.