Every word counts: A multilingual analysis of individual human alignment with model attention

Stephanie Brandl
Department of Computer Science
University of Copenhagen
brandl@di.ku.dk

Nora Hollenstein
Center for Language Technology
University of Copenhagen
nora.hollenstein@hum.ku.dk

Abstract

Human fixation patterns have been shown to correlate strongly with Transformer-based attention. Those correlation analyses are usually carried out without taking into account individual differences between participants and are mostly done on monolingual datasets making it difficult to generalise findings. In this paper, we analyse eye-tracking data from speakers of 13 different languages reading both in their native language (L1) and in English as language learners (L2). We find considerable differences between languages but also that individual reading behaviour such as skipping rate, total reading time and vocabulary knowledge (LexTALE) influence the alignment between humans and models to an extent that should be considered in future studies.

1 Introduction

Recent research has shown that relative importance metrics in neural language models correlate strongly with human attention, i.e., fixation durations extracted from eye-tracking recordings during reading (Morger et al., 2022; Eberle et al., 2022; Bensemann et al., 2022; Hollenstein and Beinborn, 2021; Sood et al., 2020). This approach serves as an interpretability tool and helps to quantify the cognitive plausibility of language models. However, what drives these correlations in terms of differences between individual readers has not been investigated.

In this short paper, we approach this by analysing (i) differences in correlation between machine attention and human relative fixation duration across languages, (ii) differences within the same language across datasets, text domains and native speakers of different languages, (iii) differences between native speakers (L1) and second language learners (L2), (iv) the influence of syntactic properties such as part-of-speech tags, and (v) the influence of individual differences in demographics, i.e., age, vocabulary knowledge, depth of processing.

Taking into account individual and subgroup differences in future research, will encourage single-subject and cross-subject evaluation scenarios which will not only improve the generalization capabilities of ML models but also allow for adaptable and personalized technologies, including applications in language learning, reading development or assistive communication technology. Additionally, understanding computational language models from the perspectives of different user groups can lead to increased fairness and transparency in NLP applications.

Contributions We quantify the individual differences in human alignment with Transformer-based attention in a correlation study where we compare relative fixation duration from native speakers of 13 different languages on the MECO corpus (Siegelman et al., 2022; Kuperman et al., 2022) to first layer attention extracted from mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021), pre-trained multilingual language models. We carry out this correlation analysis on the participants' respective native languages (L1) and data from an English experiment (L2) of the same participants. We analyse the influence of processing depth, i.e., quantifying the thoroughness of reading through the readers' skipping behaviour, part-of-speech (POS) tags, and vocabulary knowledge in the form of LexTALE scores on the correlation values. Finally, we compare correlations to data from the GECO corpus, which contains English (L1 and L2) and Dutch (L1) eye-tracking data (Cop et al., 2017).

The results show that (i) the correlation varies greatly across languages, (ii) L1 reading data correlates less with neural attention than L2 data, (iii) generally, in-depth reading leads to higher correlation than shallow processing. Our code is available at github.com/stephaniebrandl/eyetracking-subgroups.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 72–77
November 20–23, 2022.

2 Related Work

Multilingual eye-tracking Brysbaert (2019) found differences in word per minute rates during reading across different languages and proficiency levels. That eye-tracking data contains language-specific information is also concluded by Berzak et al. (2017), who showed that eye-tracking features can be used to determine a reader's native language based on English text.

Individual differences The neglection of individual differences is a well-known issue in cognitive science, which leads to theories that support a misleading picture of an idealised human cognition that is largely invariant across individuals (Levinson, 2012). Kidd et al. (2018) pointed out that the extent to which human sentence processing is affected by individual differences is most likely underestimated since psycholinguistic experiments almost exclusively focus on a homogeneous subsample of the human population (Henrich et al., 2010).

Along the same lines, when using cognitive signals in NLP, most often the data is aggregated across all participants (Hollenstein et al., 2020; Klerke and Plank, 2019). While there is some evidence showing that this leads to more robust results regarding model performance, it also disregards differences between subgroups of readers.

Eye-tracking prediction and correlation in NLP State-of-the-art word embeddings are highly correlated with eye-tracking metrics (Hollenstein et al., 2019; Salicchi et al., 2021). Hollenstein et al. (2021) showed that multilingual models can predict a range of eye-tracking features across different languages. This implies that Transformer-based language models are able to extract cognitive processing information from human signals in a supervised way. Moreover, relative importance metrics in neural language models correlate strongly with human attention, i.e., fixation durations extracted from eye-tracking recordings during reading (Morger et al., 2022; Eberle et al., 2022; Bensemann et al., 2022; Hollenstein and Beinborn, 2021; Sood et al., 2020).

3 Method

We analyse the Spearman correlation coefficients between first layer attention in a multilingual language model and relative fixation durations extracted from a large multilingual eye-tracking corpus, including 13 languages (Siegelman et al., 2022; Kuperman et al., 2022) as described below.

Total fixation time (TRT) per word is divided by the sum over all TRTs in the respective sentence to compute relative fixation duration for individual participants, similar to Hollenstein and Beinborn (2021).

We extract first layer attention for each word from mBERT[1], XLM-R[2] and mT5[3], all three are multilingual pre-trained language models. We then average across heads. We also test gradient-based saliency and attention flow, which show similar correlations but require substantially higher computational cost. This is in line with findings in Morger et al. (2022).

Eye-tracking Data The L1 part of the MECO corpus contains data from native speakers reading 12 short encyclopedic-style texts (89-120 sentences) in their own languages[4] (parallel texts and similar texts of the same topics in all languages), while the L2 part contains data from the same participants of different native languages reading 12 English texts (91 sentences, also encyclopedic-style). For each part, the complete texts were shown on multiple lines on a single screen and the participants read naturally without any time limit. Furthermore, language-specific LexTALE tests have been carried out for several languages in the L1 experiments and the English version for all participants in the L2 experiment. LexTALE is a fast and efficient test of vocabulary knowledge for medium to highly proficient speakers (Lemhöfer and Broersma, 2012).

For comparison, we also run the experiments on the GECO corpus (Cop et al., 2017), which contains eye-tracking data from English and Dutch native speakers reading an entire novel in their native language (L1, 4921/4285 sentences, respectively), as well as a part where the Dutch speakers read English text (L2, 4521 sentences). The text was presented on the screen in paragraphs for natural unpaced reading.

[1] https://huggingface.co/bert-base-multilingual-cased

[2] https://huggingface.co/xlm-roberta-base

[3] https://huggingface.co/google/mt5-base

[4] The languages in MECO L1 include: Dutch (nl), English (en), Estonian (et), Finnish (fi), German (de), Greek (el), Hebrew (he), Italian (it), Korean (ko), Norwegian (no), Russian (ru), Spanish (es) and Turkish (tr).

		MECO													GECO	
		de	el	en	es	et	fi	he	it	ko	nl	no	ru	tr	en	nl
L1	mBERT	0.45	0.57	0.27	0.42	0.52	0.51	0.49	0.35	0.45	0.38	0.41	0.53	0.48	0.26	0.26
	XLM-R	0.53	0.66	0.37	0.54	0.6	0.59	0.55	0.47	0.51	0.48	0.52	0.65	0.53	0.27	0.28
	mT5	0.31	0.45	0.11	0.24	0.37	0.36	0.27	0.16	0.35	0.27	0.23	0.3	0.23	0.16	0.23
L2	mBERT	0.32	0.33	0.26	0.32	0.32	0.32	0.33	0.34	-	0.3	0.31	0.33	0.33	-	0.29
	XLM-R	0.42	0.43	0.35	0.41	0.42	0.42	0.42	0.45	-	0.39	0.4	0.42	0.43	-	0.29
	mT5	0.11	0.13	0.08	0.12	0.13	0.13	0.12	0.13	-	0.11	0.11	0.13	0.13	-	0.18

Table 1: Spearman correlation between first layer attention and total reading time for each language and different models.[4] Correlation values are calculated individually per participant and sentence and averaged across both afterwards. First 3 rows show results for L1 languages and the remaining rows show results for the same participants on the L2 English reading task. English L2 data for Korean (ko) participants in MECO and English L2 participants in GECO is not available.

4 Results

In the following, we show results for the correlation analysis across languages and an in-depth analysis on different influences on those correlations.

Languages We compute the Spearman correlation between relative fixation and first layer attention per sentence and average across sentences for all individual participants. We show correlation values averaged across participants for each language (L1) and corresponding data for English L2 in Table 1. We can see considerable differences between the languages, particularly in L1 with higher correlation values, e.g., for mBERT (> 0.5) for *et, fi, el, ru* and lower values (< 0.4) for *nl, en, it*. Correlations for XLM-R are about 0.1 higher and for mT5 $0.1 - 0.2$ lower compared to mBERT. The correlation for English L2 are very similar between languages (0.3-0.34, mBERT) and lowest for the English L1 participants (0.26, mBERT). Correlation values for GECO are slightly lower for the Dutch experiments but in the same range for the English part.

Processing depth To further analyse the different correlation values, particularly the low correlation in the L2 experiment for English native speakers, we look into skipping rates and total reading times and hereby focus on mBERT to make results more comparable to Eberle et al. (2022). Analyses on mT5 and XLM-R show similar results. Figure 1 shows skipping rates and total reading times computed for individual participants on the entire dataset versus individual correlation values as computed above. We find significant correlations ($p < 0.01$) for both skipping rate vs. correlation values ($-0.41/-0.34$) and TRT vs. correlation values ($0.19/0.32$) for L1 and L2 respectively. This indicates that more thorough reading, i.e., less skipping and more time per word, leads to higher correlation with first layer attention. We also see those correlations at language-level for some languages where *he, fi, ru* show highest scores at $-0.7, -0.63, -0.59$, respectively. For GECO, we find similar trends for English (L1 and L2) but not for Dutch.

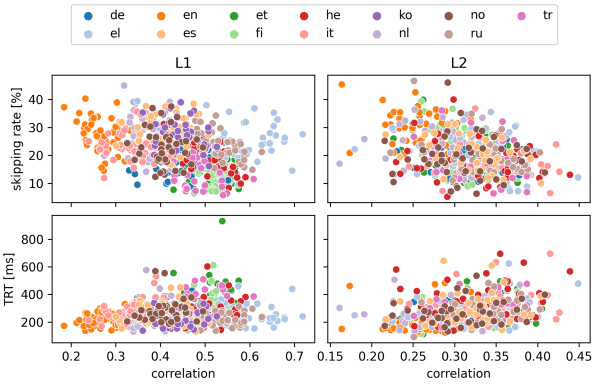

Figure 1: Correlation values for individual participants versus skipping rate (upper) and total reading time (lower) for L1 (left) and L2 (right) data. Spearman correlation was calculated on sentence-level and then averaged. Results are shown for mBERT.

POS We look deeper into cross-lingual differences and show correlation values on token-level for 6 frequent POS tags in Figure 2. We extract relative fixations, standardise them to mean=0 and std=1 and average them across participants before computing the Spearman correlation with first layer attention values. We use POS-tagging models from *spacy* and show results for the languages where

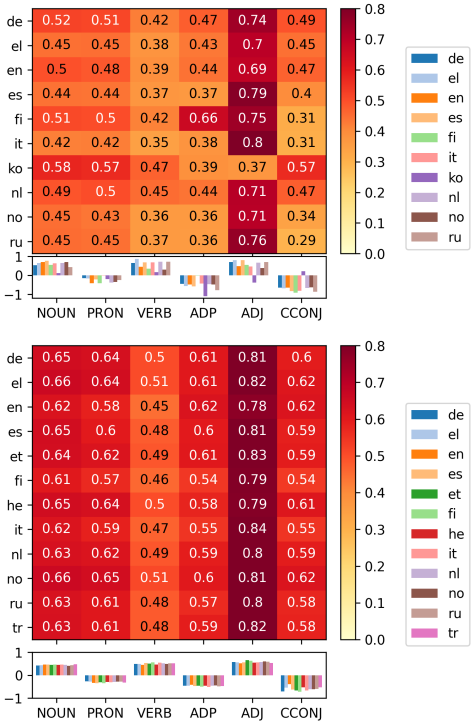

Figure 2: Spearman correlations between human fixation and different languages for L1 (*upper*) and L2 (*lower*) for selected POS tags. Barplots show average attention value after standardisation (mean=0, std=1) for respective POS tag and model. For L1 only those languages are presented with an available POS-tagging model. Note that correlations are computed at token-level (not at sentence-level) which might cause higher correlations in L2. Results are shown for mBERT.

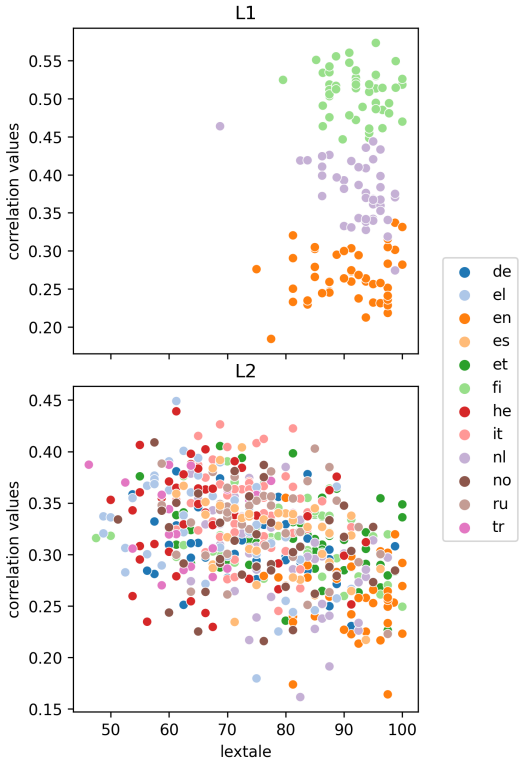

Figure 3: Spearman correlation values versus LexTALE score for individual participants for selected languages in L1 (*en*, *nl* and *fi*) and all speakers in L2. Values for *fi* in L1 were rescaled (with 100/88) to make them comparable. Results are shown for mBERT.

respective models are available.[5] Correlations for L1 are distributed similarly across different POS tags where adjectives show the highest correlation whereas verbs, although they carry an important part of the fixations, correlate much less. Only *Korean* poses an exception here where adjectives do not play the most prominent role in human attention and also correlate much less. Here, nouns, pronouns, verbs and coordinating conjunctions correlate higher than in any other language and also much higher than adjectives. More research is required to interpret this finding. For L2, we see a very homogeneous distribution between languages and a similar distribution across POS tags as in most L1 experiments.

LexTALE We show LexTALE scores for *English* L2 and *fi, en, nl* for L1 versus correlation values in Figure 3. We find a negative correlation for Dutch speakers in L1 -0.36 and for the entire L2 data of -0.42 ($p < 0.05$) suggesting that higher LexTALE scores lead to lower correlation with first layer attention.

5 Discussion & Conclusion

Our results show that the correlation between relative fixation duration and first layer attention varies greatly across languages when read by native speakers. These differences can be attributed in part to the depth of processing: Languages such as Finnish and Greek, which show high total reading times, show a more evenly distributed correlation pattern across the most frequent parts of speech. Moreover, L1 English shows a high skipping rate and the lowest correlations. We find that more careful in-depth reading – processing more words for a longer time – correlates more strongly with attention than fast shallow reading. This is in line with previous research showing that attention patterns in BERT carry high entropy values, i.e., are broadly distributed, particularly in the first layers (Clark et al., 2019), which also leads to higher correlation with fixation duration (Eberle et al., 2022).

[5] https://spacy.io/usage/models

The differences in skipping rate have various origins. On one hand, skipping rate is regulated by word length (Drieghe et al., 2004), which explains the lower skipping rate of agglutinative languages such as Finnish and Turkish (Siegelman et al., 2022), and in turn their higher correlation to mBERT attention. On the other hand, word skipping is affected by L2 reading proficiency. More skilled learners make fewer fixations and skip more words (Dolgunsöz and Sarıçoban, 2016). This is reinforced by our comparison between English L2 and native English reading (which shows lower correlation). This finding is also supported by our analysis on the LexTALE vocabulary test. LexTALE accurately estimates proficiency even at high levels (Ferré and Brysbaert, 2017). Our results show that higher test scores lead to lower correlation with attention. Again, this is due to the reading depth: highly proficient readers have a higher skipping rate (Eskenazi and Folk, 2015).

We furthermore looked at the influence of age and gender but could not find any meaningful differences. This might be due to the fact that all participants were university students, most of them under the age of 30, thus representing a very specific group of the overall population. It is also important to note that most of the languages in MECO are Indo-European and only 4 are not using the Latin script.

In summary, we have shown the impact of various subgroup characteristics reflected in reading and how they affect the correlation to neural attention. We argue that these differences should be taken into account when leveraging human language processing signals for NLP.

Acknowledgements

We thank Daniel Hershcovich for proof-reading and valuable inputs on the manuscript. SB was partially funded by the Platform Intelligence in News project, which is supported by Innovation Fund Denmark via the Grand Solutions program and by the European Union under the Grant Agreement no. 10106555, FairER. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor REA can be held responsible for them.

References

Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87, Dublin, Ireland. Association for Computational Linguistics.

Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting native language from gaze. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 541–551, Vancouver, Canada. Association for Computational Linguistics.

Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of memory and language*, 109:104047.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emrah Dolgunsöz and Arif Sarıçoban. 2016. Word Skipping in Reading English as a Foreign Language: Evidence from Eye Tracking. *East European Journal of Psycholinguistics*.

Denis Drieghe, Marc Brysbaert, Timothy Desmet, and Constantijn De Baecke. 2004. Word skipping in reading: On the interplay of linguistic and visual factors. *European Journal of Cognitive Psychology*, 16(1-2):79–103.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.

Michael A Eskenazi and Jocelyn R Folk. 2015. Reading skill and word skipping: Implications for visual and linguistic accounts of word skipping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(6):1923.

Pilar Ferré and Marc Brysbaert. 2017. Can LexTALE-Esp discriminate between groups of highly proficient Catalan–Spanish bilinguals with different language dominances? *Behavior research methods*, 49(2):717–723.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Nora Hollenstein and Lisa Beinborn. 2021. Relative importance in sentence processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150, Online. Association for Computational Linguistics.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 538–549, Hong Kong, China. Association for Computational Linguistics.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Evan Kidd, Seamus Donnelly, and Morten H Christiansen. 2018. Individual differences in language acquisition and processing. *Trends in cognitive sciences*, 22(2):154–169.

Sigrid Klerke and Barbara Plank. 2019. At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, pages 1–35.

Kristin Lemhöfer and Mirjam Broersma. 2012. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, 44(2):325–343.

Stephen C Levinson. 2012. The original sin of cognitive science. *Topics in cognitive science*, 4(3):396–403.

Felix Morger, Stephanie Brandl, Lisa Beinborn, and Nora Hollenstein. 2022. A cross-lingual comparison of human and model relative word importance. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 11–23, Gothenburg, Sweden. Association for Computational Linguistics.

Lavinia Salicchi, Alessandro Lenci, and Emmanuele Chersoni. 2021. Looking for a role for word embeddings in eye-tracking features prediction: Does semantic similarity help? In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 87–92, Groningen, The Netherlands (online). Association for Computational Linguistics.

Noam Siegelman, Sascha Schroeder, Cengiz Acarturk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, and Victor Kuperman. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior Research Methods*, pages 1–21.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.