A Cross-lingual Comparison of Human and Model Relative Word Importance

Felix Morger[*]
Gothenburg University

Stephanie Brandl
University of Copenhagen

Lisa Beinborn
Vrije Universiteit Amsterdam

Nora Hollenstein
University of Copenhagen

Abstract

Relative word importance is a key metric for natural language processing. In this work, we compare human and model relative word importance to investigate if pretrained neural language models focus on the same words as humans cross-lingually. We perform an extensive study using several importance metrics (gradient-based saliency and attention-based) in monolingual and multilingual models, including eye-tracking corpora from four languages (German, Dutch, English, and Russian). We find that gradient-based saliency, first-layer attention, and attention flow correlate strongly with human eye-tracking data across all four languages. We further analyze the role of word length and word frequency in determining relative importance and find that it strongly correlates with length and frequency, however, the mechanisms behind these non-linear relations remain elusive. We obtain a cross-lingual approximation of the similarity between human and computational language processing and insights into the usability of several importance metrics.

1 Introduction

Large pretrained neural language models, such as BERT (Devlin et al., 2019), have in recent years demonstrated performance equal to that of humans in a range of natural language understanding tasks (Wang et al., 2019). This begs the question of whether the processing and encoding of these models reflect language properties as described by language experts, such as in grammar, semantics, pragmatics and logic and, furthermore, whether the models process language similarly to humans. While extensive research is being done to answer this question, such as inquiries into what linguistic knowledge is encoded into contextual word representations (Clark et al., 2019; Vulić et al., 2020), how linguistic information is processed (Tenney et al., 2019) and the effects of architectural choices (Rogers et al., 2020), more recent research inspired by psycholinguistics has emerged, which directly compares cognitive signals of language processing to pretrained language models. By using tools such as eye-tracking features and brain activity data (Abdou, 2022; Goldstein et al., 2022; Hollenstein et al., 2020a), this line of research skips the step of having to collect human judgments from speech or text data by directly comparing them to sources of cognitive data. As such, this approach is a direct means of testing whether models process language similarly to humans or, in other words, of evaluating the *cognitive plausibility* of computational language processing.

One method leveraging cognitive data has been to extract relative word importance, a key metric for natural language processing, from eye-tracking data in order to compare these to relative word importance extracted from state-of-the-art pretrained language models. This has been studied for normal English reading (Hollenstein and Beinborn, 2021; Bensemann et al., 2022), in task-specific reading (Eberle et al., 2022), and in question answering settings (Sood et al., 2020). In this work, we continue in this line of research but apply it across several languages to measure the extent to which pretrained language models focus on the same words as humans cross-lingually. We obtain *human relative word importance* from eye-tracking data (total reading time) and *model relative word importance* from pretrained language models using saliency and attention-based methods. These methods have in recent years been developed for the purpose of explainability, however, which methods serve best for this purpose has been a point of contention (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).

The goal of this study is two-fold: On the one hand, we aim to obtain a rough estimate of the similarity between human and computational natural

[*] E-mail: felix.morger@gu.se

Proceedings of the 2022 CLASP Conference on (Dis)embodiment, pages 11–23
September 15–16, 2022. ©2022 Association for Computational Linguistics

language processing and, on the other hand, from a usability point of view, see which importance methods best approximate human relative word importance. We compare four methods for calculating relative word importance in pretrained language models, namely first-layer attention, last-layer attention, attention flow, and gradient-based saliency. To investigate whether the same trends hold across multiple languages and are not particular artifacts of one language, we use eye-tracking corpora of four different languages (English, Dutch, German, and Russian) and compare both monolingual and multilingual language models. More precisely, we make this comparison by looking at how human and model relative word importance statistically correlate across different languages.

Lexical properties such as word frequency and length are known to have a large effect on eye movements of any language (Just and Carpenter, 1980; Levy, 2008). Therefore, in an additional investigation, we analyze their impact on the fit between human and model relative importance.

To sum up, this work examines the following research questions:

Q1: Do human and model relative word importance correlate across languages?

Q2: Is there a difference between language-specific and multilingual language models?

Q3: Is there a difference between gradient-based saliency and the attention-based methods first-layer attention, last-layer attention, and attention flow?

Q4: To what extent is human and model relative word importance relying on word length and word frequency?

Contributions We show that human and model word importance correlate strongly in varying degrees across languages (English, Dutch, German and Russian), although the observed differences appear to be more corpus-specific than language-specific (Q1). We observe a slightly stronger correlation of monolingual models over multilingual models, in particular for first-layer attention and attention flow (Q2). We see that other attention-based methods than last-layer attention, i.e., first-layer and attention flow correlate strongly to human eye-tracking data with attention flow being on par with saliency for monolingual models (Q3). We see a strong correlation with the baselines (positive correlation to word length and negative correlation to word frequency). When using linear regression analysis to measure the ability of word length, word frequency and model relative word importance to predict human relative word importance, we see that word frequency and word length increase the predictive power over model relative word importance alone, indicating that these baselines are not sufficiently accounted for by the models (Q4). The code for our experiments is available online.[1]

2 Related Work

This work lies at the intersection of psycholinguistics, interpretability of neural networks, and natural language processing. More specifically, there are two current streams of research that this study directly draws from, namely relative importance metrics and cognitive analysis of natural language processing. Below, we outline the related works in these two subfields.

2.1 Relative Importance Metrics

Approaches for extracting relative word importance of Transformer-based models can be grouped into gradient-based, propagation-based, occlusion-based, and attention-based methods (Bastings and Filippova, 2020, Section 3). In this work, we focus on attention-based and gradient-based methods (see Section 4).

Attention is a key component of Transformer models and multiple studies have analyzed how attention weights are distributed across tokens. It has, for example, been shown that attention at different layers in Transformer models targets different linguistic aspects. For instance, Vig and Belinkov (2019) find that attention in a GPT-2 model targets different parts of speech and depths of dependency relations at different layers within the model and Li et al. (2021) show that different layers of transformer language models perform best when detecting different types of linguistic anomalies. Also, the findings by Tenney et al. (2019) indicate that in BERT earlier layers encode more word-level information than later layers when comparing performance across different language-level tasks from part-of-speech tagging to anaphora resolution.

The methodological merit of attention weights as a measure of relative importance has, however, been questioned. For one, the calculated attention attends to input representations, not the input itself, and these representations can mix in information from other inputs, thus diluting the relative impor-

[1] https://github.com/felixhultin/relative_importance.

tance strength of the original input token (Bastings and Filippova, 2020). Moreover, different attention distributions can lead to the same predictions, making the relative importance of attention weights ambiguous (Jain and Wallace, 2019). To address the unreliability of attention weights, Abnar and Zuidema (2020) propose *attention flow*, a mechanism which computes maximum flow values, from hidden embeddings to input tokens.

2.2 Cognitive Analysis of Natural Language Processing

Using cognitive data to evaluate NLP has emerged as a novel method for interpreting NLP systems (Toneva and Wehbe, 2019; Ettinger, 2020; Hollenstein et al., 2019). The motivation behind this research is to assess whether models encode, process, or output language similarly to humans and, thus, provide measurements of their cognitive plausibility (Keller, 2010).

In recent years, more eye-tracking corpora from natural reading have become available in multiple languages (see section 3.1). Although cognitive data, including eye-tracking corpora, have been available as digitized formats for a long time, only recently have they been methodically deployed for the cognitive analysis of NLP systems. For example, the CMCL shared evaluation task uses ZuCo for the modeling of eye-tracking features. In this task, language models, such as BERT, are used to predict eye-tracking features (number of fixations, first fixation duration, total reading time, etc.) (Hollenstein et al., 2021). This work is similar to ours, but instead of fine-tuning the model to predict eye-tracking features, we see if the relative word importance as extracted by different methods correlates to mean total reading time.

Further work using other sources than eye-tracking corpora is for example Ettinger (2020), who proposed a psycholinguistic test suite to diagnose language models' predictions in context using electroencephalogram (EEG). Moreover, Abnar et al. (2019) use functional magnetic resonance imaging (fMRI) and representational similarity analysis (RSA) to compare representations of the brain and pretrained language models.

In terms of using relative importance metrics, previous studies have shown that attention weights do not correspond to human relative word importance. For example, Sood et al. (2020) compare attention weights from the last layer of Transformer models to human gaze data. They show that a higher correlation between model attention and human attention does not necessarily yield better performance in downstream NLP tasks. Hollenstein and Beinborn (2021) also find that attention has a weak correlation to human gaze data. Recently, Eberle et al. (2022) have found that attention flow from transformer models correlates strongly with human fixation times in task-specific English reading.

Finally, gradient-based methods have been proposed as a better method than attention weights at approximating the relative importance of input words in neural networks (Bastings and Filippova, 2020). Hollenstein and Beinborn (2021) additionally show that gradient-based saliency might be a cognitively more plausible interpretability metric than attention weights.

We follow these results and provide a large cross-lingual comparison of human eye-tracking data to a range of relative importance metrics, including gradient-based saliency, first and last-layer attention, and attention flow.

3 Data

3.1 Eye-tracking Corpora

We use eye-tracking data collected from native readers of the following corpora to extract the human relative importance metrics based on the mean total reading time of each word (see Table 1). For English, we use the GECO corpus, which contains eye tracking data from English monolinguals reading an entire novel (Cop et al., 2017), and the ZuCo corpus (Hollenstein et al., 2018, 2020b), which includes eye-tracking data of full sentences from movie reviews and Wikipedia articles.[2] For Dutch, we also use the GECO corpus, which additionally contains eye tracking data from Dutch readers that were presented with the same novel in their native language (Cop et al., 2017). For German, we leverage the Potsdam Textbook Corpus, which contains 12 short passages from college-level biology and physics textbooks, which are read by expert and laymen German native speakers (Jäger et al., 2021). We also use the Russian Sentence Corpus which includes naturally occurring sentences extracted from the Russian National Corpus (Laurinavichyute et al., 2019).[3] We exclude a small set

[2] We use Tasks 1 and 2 from ZuCo 1.0 and Task 1 from ZuCo 2.0.

[3] https://ruscorpora.ru

Language	Corpus	Subjs.	Sents.	Sent. length	Tokens	Types	Word length
English	GECO	14	4,559	10.5 (1–69)	56,410	5,916	4.6 (1–33)
	ZuCo	30	853	19.5 (1–68)	20,545	5,560	5.0 (1–29)
Dutch	GECO	19	4,863	11.6 (1–60)	59,716	5,575	4.5 (1–22)
German	PoTeC	75	89	19.5 (5–51)	1,895	847	6.5 (2–33)
Russian	RSC	103	143	9.4 (5–13)	1,357	993	5.7 (1–18)

Table 1: Descriptive statistics of all eye-tracking datasets. Sentence length and word length are expressed as the mean with the min-max range in parentheses. **Sents.** is the number of the subset of sentences we process for this work, while sentence length and word length are calculated from all sentences in the corpora.

of sentences from the original corpora because of token alignment issues.

3.2 Language Models

All monolingual models and the multilingual model are based on the BERT architecture (Devlin et al., 2019). We use the pretrained checkpoints from the HuggingFace repository of the multilingual base model and language-specific monolingual base models. See Table 3 in the Appendix for the complete list of models and references.

4 Method

For each sentence in the eye-tracking corpora, we calculate human and model relative word importance values. The same sentences are, however, tokenized differently by the built-in tokenizers of the pretrained language models, resulting in longer sequences of relative word importance values than those obtained from humans. To remedy this, we align human and model importance values by discarding the importance values of special tokens (e.g. [SEP] and [CLS]) and merging subtokens and adding their values. Once aligned, we calculate Spearman's correlation coefficient ρ between human and model relative word importance for each sentence. Finally, we calculate an average Spearman's ρ across all sentences for each eye-tracking corpus (human relative word importance) and model, corpus, and importance metric tuple (model relative word importance). Additionally, using the same procedure, we explore the correlation of human and model relative word importance to word length and frequency baselines.

We analyze the following importance metrics:

Human relative importance In this work, we use the *total reading time* per word in the eye tracking corpora as the source for defining human relative word importance. It refers to the sum of all fixation durations for each word including regressions (i.e. when a subject goes back to the same word after the first pass). We use the average total reading time across all subjects and normalize the resulting values such that each word is assigned an importance value between 0 and 1, and all values within a sentence sum up to 1. These values are calculated sentence by sentence.

Gradient-based saliency As described in Hollenstein and Beinborn (2021), we define a saliency vector for a masked token to indicate the importance of each of the tokens in the context of correctly predicting the masked token (Madsen, 2019). The saliency s_{ij} for input token $\mathbf{x}_j$ for the prediction of the correct token $\mathbf{t}_i$ is calculated as the Euclidean norm of the gradient of the logit for x_i:

$$s_{ij} = \|\nabla_{\mathbf{x}_j} f_{t_i}(\mathbf{X}_i)\|_2 \qquad (1)$$

Last-layer attention We approximate relative importance using the attention values from the last layer and calculate the mean of all heads of each Transformer model as Sood et al. (2020).

First-layer attention Previous work has indicated that earlier layers encode information closer to word-level than later layers (Tenney et al., 2019). Therefore, we also include the first-layer attention weights by averaging over all heads to approximate relative word importance.

Attention flow Finally, we compute attention flow (Abnar and Zuidema, 2020), which has been shown to correlate stronger with human gaze than raw attention weights (Eberle et al., 2022). Attention flow considers the attention graph as a flow network and computes maximum flow values from later attention layers to the input embedding layer. Unlike raw attention weights, which consider token importance at layers in isolation, attention flow computes importance scores that account for mixing of information across layers and, thus, identifies

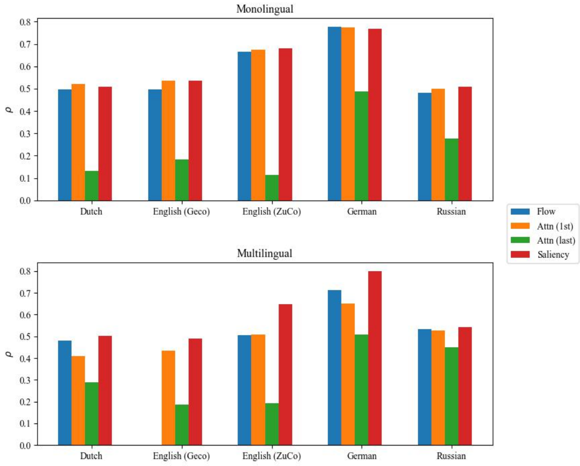

Figure 1: Upper: Spearman's correlation ρ between human (total reading time) and model relative word importance (BERT monolingual models). Lower: Spearman's correlation ρ between human (total reading time) and model relative word importance (BERT multilingual model).

the important tokens for the model prediction. Because of the high computational resources needed to calculate attention flow, we do not calculate attention flow of the BERT multilingual model for English (GECO).

Baselines We compare human and model relative word importance to two word-level baselines: word length (the number of characters in a word) and word frequency (the proportion of times a word occurs in a corpus). To obtain the word frequencies, we use the wordfreq Python package (Speer et al., 2018) (version 2.3.2), which calculates token frequencies based on corpora from different Internet text resources, such as Wikipedia, Google Books, and Reddit.

Regression Analysis In addition, we use a mixed linear regression analysis (ordinary least squares) to measure the extent to which, model relative word importance, word frequency, and word length can predict human relative word importance. We let human relative word importance be the dependent variable and fit multiple linear regression models with different combinations of model word importance, word frequency and word length as independent variables. This is done to measure each and every variable's effect in isolation. We analyze the resulting coefficient of determination R^2.

In the Spearman correlation analysis outlined above, the correlations were calculated per sentence and then averaged. In contrast, we now fit the model to tokens which means that all relative word importance values and all word lengths and frequencies are fitted into the same model.[4].

Since all independent variables (word frequency, word length and model relative word importance) and the dependent variable (human relative word importance) are intrinsically skewed, we log-transform all data. Furthermore, we use an extended version of linear regression (mixed linear regression (Gałecki and Burzykowski, 2013)) to deal with dependency between samples (i.e., one word appearing more than twice) which otherwise would break the assumption of linear regression models that each observation is independent of each other.

5 Results

5.1 Human vs. Model Word Importance

Figure 1 shows the Spearman correlation between human relative word importance and model relative word importance of importance methods for each eye-tracking corpus. The results show a strong correlation (ρ > .5) between human and model word importance across all languages. There are, however, considerable differences between languages. For example, German reaches a Spearman's ρ of .8, while Russian, English (GECO) and Dutch (GECO) only reach .5 (Q1). When comparing the multilingual BERT model to language-specific BERT models, we observe for some importance metrics, attention flow and first-layer attention, a slightly stronger correlation to monolingual models. In the German and English (ZuCo) monolingual models, in particular, first-layer attention and flow are equally strong as saliency while in the multilingual model they all have more than .1 weaker correlation. Attention first-layer seems even more strongly correlated to monolingual models, where we see +.11 and +.17 for the language-specific BERT model of Dutch (GECO), English (GECO) and ZuCo (English), respectively. Russian, however, is a slight outlier in that it has a +.3 difference in favor of the multilingual BERT model (Q2).

When comparing importance methods, we see similar results to previous findings on English data. Saliency shows a strong correlation to human relative word importance, while last-layer attention shows a weaker correlation. Furthermore, attention flow and, surprisingly first-layer attention in most

[4] Most sentences are too short for the number of independent variables we use to fit a linear regression model.

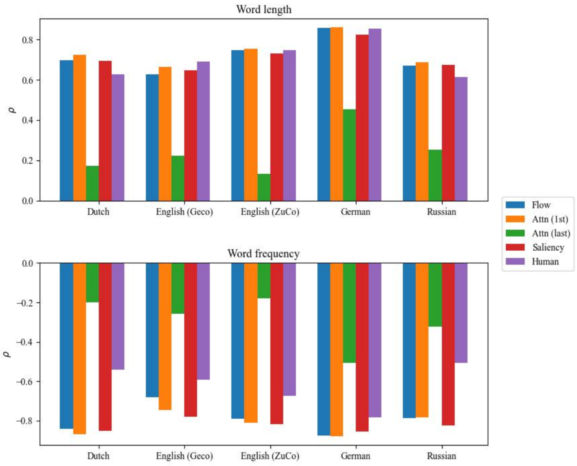

Figure 2: Upper: Spearman's correlation ρ between word length and human (total reading time) and model (BERT monolingual) relative word importance. Lower: Spearman's correlation ρ between word length and human (total reading time) and model (BERT monolingual) relative word importance.

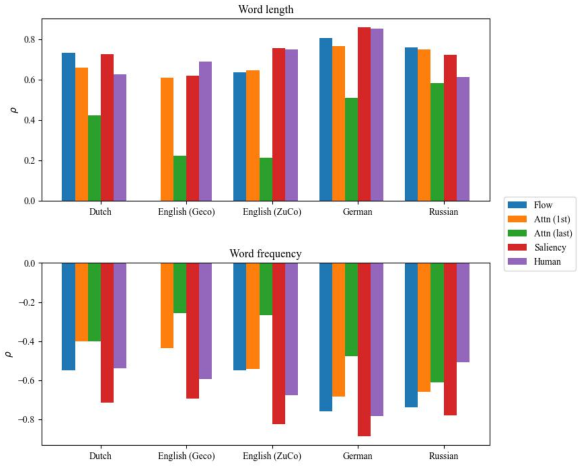

Figure 3: Upper: Spearman's correlation ρ between word length and human (total reading time) and model (BERT multilingual) relative word importance . Lower: Spearman's correlation ρ between word length and human (total reading time) and model (BERT multilingual) relative word importance.

cases, show similar strength than saliency, albeit slightly weaker for multilingual models (Q3).

Finally, though not specifically defined in the goals and research questions of this study, we make the separate, but important, observation that the most impactful variable for correlation strength seems to be the size and text domain of the eye-tracking corpora. This becomes apparent when comparing Dutch (GECO), English (GECO) and English (ZuCo). Even though English (GECO) and English (ZuCo) are of the same language, the performance on English (GECO) and Dutch (GECO) are quite similar, while not very similar to English (ZuCo). While the language-specific impact on the results is difficult to grasp due to the differences between the eye-tracking corpora, we, nonetheless, see the same trends hold for all four languages.

5.2 Corpus Statistical Baselines

Figures 2 and 3 show the correlation of word frequency and length baselines to human and model relative word importance. We see a strong correlation between models of all languages and the two baselines word length and word frequency: A strong *positive* correlation to word length and a strong *negative* correlation to word frequency, as also observed by Hollenstein and Beinborn (2021).

For the baselines, however, we see considerable differences between languages. German shows the strongest correlation for word length with 0.85 for humans and 0.82 and 0.85 for mono- and multilingual BERT, respectively, and also the strongest negative correlation for word frequency with -0.78 for humans and -0.85 as well as -0.87 for mono- and multilingual BERT, respectively. Russian shows the weakest human correlation to the baselines, 0.61 for word length and -0.51 for word frequency, while having a relatively strong saliency baseline correlation of 0.72 and 0.67 for word length as well as -0.78 and -0.82 for word frequency.

Looking at the relative word importance metrics, which had the strongest correlation to human importance, namely saliency and attention flow, we see that their correlation strength with respect to the baselines correlate equally strong or stronger than their human counterparts. This indicates that the more similar their baselines are to the human baselines, the stronger they correlate in terms of relative word importance (see previous section). This is especially the case when looking at (1) attention last-layer, where there weaker correlation to human relative word importance is also reflected in its weaker correlation to word frequency and word length, which are much lower than its human counterpart and (2) word frequency, where the model relative word importance of Dutch (GECO), English (GECO) and Russian have a weaker correlation to human relative word importance but a considerable stronger negative correlation to word frequency than human relative word importance

Table 2: Linear regression (R^2) fitted to predict human relative word importance from word frequency and/or word length.

	freq	length	freq+length
Dutch	0.08	0.15	0.16
English (GECO)	0.07	0.12	0.14
German	0.31	0.38	0.41
Russian	0.22	0.49	0.49
English (ZuCo)	0.13	0.26	0.30

has to word frequency. This effect does not, however, seem to be as pronounced with word length.

These results show that word length and word frequency are powerful indicators of word importance and support the presumption that they play an important role in determining the correlation strength between human and model relative word importance (Q4). In the next subsection, we will try and quantify how much these baselines account for this relation, by measuring the explanatory power word length, word frequency and model relative word importance have in predicting human relative word importance.

5.3 Word Length & Frequency Regression Analysis

We fitted linear mixed models to predict human word importance using either word frequency, word length, model relative word importance, or combinations of the features. Table 2 shows the R^2 results for using word frequency and word length as independent variables, and Figure 4 shows the results for using model relative word importance as an independent variable in combination with word frequency and word length. See Figure 5 and Table 4 in the Appendix for full results.

In Table 2 we see a weak R^2 score or in other words a weak *linear* relationship between human relative word importance and word frequency and word length. Word frequency, however, appears to have a weaker R^2 than word length and differences are large between corpora. Russian, for example, has four times stronger R^2 for word length than English (GECO). Using both word frequency and word length (freq+length) appears only to be as strong as the strongest word length value (length), such that a combination of word frequency and word length (freq+length) does not make the relationship stronger.

Comparing Figure 4 to Table 2 we see a much stronger linear relationship (R^2) when model relative word importance is used as an independent variable. When combined with word frequency (model+freq) we see a considerable increase of R^2, but combined with word length (model+length and model+freq+length) we see an even stronger linear relationship. Similarly to Table 2, combining word frequency and word length (freq+length) only gives as much benefit as adding length to model word importance (model+length and model+freq+length).

Comparing the R^2 of model importance, we see different scores than that of the results in section 5.1 and section 5.2. Here, we see saliency achieving the lowest R^2 across all models and corpora, meanwhile the attention-based metrics (attention first/last layer and flow) show a much larger R^2. Although comparing the R^2 and Spearman's ρ is not equal due to the methodological differences outlined in Section 4, this difference nevertheless suggest that relation between saliency and human relative word importance is less linear in nature than attention-based ones.

6 Discussion

6.1 Findings on Human vs. Model Relative Word Importance

This cross-lingual study shows that model relative word importance has a strong correlation to human relative word importance. We confirm the findings of other English-based studies that saliency (Hollenstein and Beinborn, 2021), first-layer attention (Bensemann et al., 2022) and attention flow (Eberle et al., 2022) show a strong correlation to human relative word importance as well as last-layer attention showing a weak correlation (Q1 & Q3). This research, thus, from a usability perspective supports the critique against using attention weights for explanation, while providing supporting evidence for the use of attention flow. Comparing monolingual and multilingual models, we see slightly stronger results for monolingual models, in particular for attention first-layer and attention flow, indicating that some importance-bearing information are more readily available in the attention weights of monolingual models. However, given that these results are not vastly different, there is still a strong argument for training multilingual models over monolingual models because of their resource-efficiency and saving of computational resources (Q2).

The secondary finding of this study that corpus-specific differences have a big impact on correlation strength, indicates the need to control the size

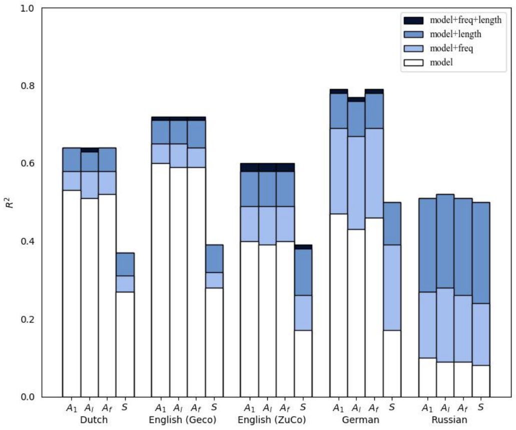

Figure 4: Linear regression (R^2) fitted to predict human relative word importance from model word importance (model), word frequency (freq), word length (length) or combinations thereof (+) (BERT monolingual). A_1, A_l, A_f, and S are short for first-layer attention, last-layer attention, attention flow, and saliency, respectively.

and text domain for cross-lingual comparison. A promising avenue for future work could be to apply our analyses to the recent multilingual eye-tracking corpus MECO (Siegelman et al., 2022).

6.2 The Role of Word Length And Word Frequency

Measuring the effect of word length and word frequency using Spearman's ρ and linear regression we find that they play an important role in determining relative word importance. First, we see that similarity in correlation strength to the word length and especially word frequency baseline mirrors a stronger correlation between model and human relative word importance (Q4). Secondly, we see that when using linear regression to quantify the linear relationships between the two lexical baselines, model relative word importance, and human relative word importance, stronger predictions can be made when model relative word importance is combined with word frequency or word length, especially the latter (see Figure 4). This suggests that word frequency and especially word length might not be sufficiently accounted for by the language models. Furthermore, the discrepancy between the lower impact of word frequency and the higher impact of word length has several potential explanations. Firstly, word frequency might be better approximated by the model than word length (which is supported by the fact that word length is not explicitly processed in Transformer-based architectures). Secondly, the relationship between word frequency and relative word importance is probably less linear than that of word length and relative word importance (as suggested by the results in Table 2) and could, therefore, not be as adequately fitted by linear regression. The nature of the relationships between word length, word frequency, and human relative word importance, thus, remains elusive. To gain clarity on this, future work could control for word length and frequency more explicitly, by, for example, grouping and comparing relative word importance by length and frequency in isolation as well as using probing tasks to test the extent to which contextual word representations themselves can predict word length and frequency.

7 Conclusion

In this work, we show that the strong correlation between relative word importance of neural language models and humans holds across several languages, namely, English, German, Dutch, and Russian. This is the case for both monolingual as well as multilingual pretrained Transformer models, which yield similar performance in our correlation analyses.

We also find that several relative importance metrics for pretrained language models, both first-layer attention and attention flow as well as saliency, perform similarly well and that these importance values, as their human counterparts, strongly correlate to word length and word frequency. However, as expected, we have found that last-layer attention correlates more weakly.

Comparing the correlations of relative word importance is a simple, easily interpretable metric for evaluating the similarity of human and computational language processing. Using this metric, we can evaluate the extent to which the model's attention compares to approximate human language processing and, thus, get a gauge of their cognitive plausibility. In addition to the BERT-based architectures we studied, looking at more recent cross-lingual models such as GPT, T5 and XLNet as well as multimodal language models would give further insights into the role of pre-training tasks as well as non-textual modalities.

Acknowledgements

We thank Alexander Koplenig for his assistance in assembling the results for the regression analysis.

References

Mostafa Abdou. 2022. Connecting neural response measurements & computational models of language: a non-comprehensive guide. *arXiv preprint arXiv:2203.05300*.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155.

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.

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.

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.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch BERT model. *CoRR*, abs/1912.09582.

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.

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.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Andrzej Gałecki and Tomasz Burzykowski. 2013. *Linear Mixed-Effects Model*, pages 245–273. Springer New York, New York, NY.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020a. 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, Emmanuele Chersoni, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. Cmcl 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23nd Conference on Computational Natural Language Learning*.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020b. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 138–146.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings* *of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Lena Jäger, Thomas Kern, and Patrick Haller. 2021. Potsdam Textbook Corpus: Potsdam textbook corpus (potec): Eye tracking data from experts and non-experts reading scientific texts. available on OSF, DOI 10.17605/OSF.IO/DN5HP.

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.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329.

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 60–67.

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

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

AK Laurinavichyute, Irina A Sekerina, SV Alexeeva, and KA Bagdasaryan. 2019. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Cyrillic. *Behavior research methods*, 51(3):1161–1178.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is bert surprised? layerwise detection of linguistic anomalies. *arXiv preprint arXiv:2105.07452*.

Andreas Madsen. 2019. Visualizing memorization in rnns. *Distill*. Https://distill.pub/2019/memorization-in-rnns.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 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.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2. 2. *Zenodo [Computer Software]*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Sarah Wiegreffe 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.

Appendix

A Limitations

The results and conclusions of this paper should be read with the following limitations in mind: (1) Differences between human subjects can be quite significant, thus, the results will also reflect this uncertainty (Kidd et al., 2018). (2) Comparing model relative word importance of saliency-based and attention-based methods to that of human relative word importance only reflects these methods' ability to mimic human behavior, but does not say anything about their ability to accurately represent the inner workings, i.e., the *faithfulness* of pretrained language models (Jacovi and Goldberg, 2020).

Table 3: List of models used for each corpora. Hugging Face path refers to the model path used to identify the model in Hugging Face repository. For models without explicit paper reference, we refer to the Hugging Face website.

Corpus	Model name	Hugging Face path	Reference
GECO (en)	BERT	bert-base-uncased	Devlin et al. (2019)
GECO (en)	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)
GECO (nl)	BERT	GroNLP/bert-base-dutch-cased	de Vries et al. (2019)
GECO (nl)	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)
ZuCo	BERT	bert-base-uncased	Devlin et al. (2019)
ZuCo	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)
Potsdam	BERT	dbmdz/bert-base-german-uncased	`https://huggingface.co/dbmdz/bert-base-german-uncased` (accessed 2022-03-15)
Potsdam	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)
Russsent	BERT	DeepPavlov/rubert-base-cased	Kuratov and Arkhipov (2019)
Russsent	BERT Multilingual	bert-base-multilingual-cased	Devlin et al. (2019)

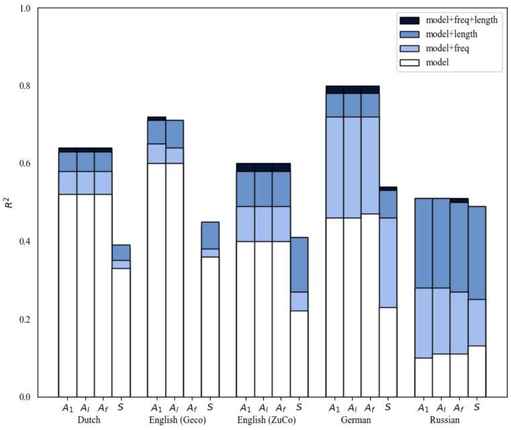

Figure 5: Linear regression (R^2) fitted to predict human relative word importance from model word importance (model), word frequency (freq), word length (length) or combinations thereof (+) (BERT multilingual). A_1, A_l, A_f, and S are short for attention (first-layer), attention (last-layer), attention flow, and saliency, respectively.

Table 4: Linear regression models R^2 measuring impact of word length, word frequency and word importance.

Language	Model	Importance	model	model+freq	model+length	model+freq+length
Dutch	BERT	Attn (1st)	0.53	0.58	0.64	0.64
		Attn (last)	0.51	0.58	0.63	0.64
		Flow	0.52	0.58	0.64	0.64
		Saliency	0.27	0.31	0.37	0.37
	mBERt	Attn (1st)	0.52	0.58	0.63	0.64
		Attn (last)	0.52	0.58	0.63	0.64
		Flow	0.52	0.58	0.63	0.64
		Saliency	0.33	0.35	0.39	0.38
English (Geco)	BERT	Attn (1st)	0.6	0.65	0.71	0.72
		Attn (last)	0.59	0.65	0.71	0.72
		Flow	0.59	0.64	0.71	0.72
		Saliency	0.28	0.32	0.39	0.39
	mBERT	Attn (1st)	0.6	0.65	0.71	0.72
		Attn (last)	0.6	0.64	0.71	0.71
		Saliency	0.36	0.38	0.45	0.44
English (ZuCo)	BERT	Attn (1st)	0.4	0.49	0.58	0.6
		Attn (last)	0.39	0.49	0.58	0.6
		Flow	0.4	0.49	0.58	0.6
		Saliency	0.17	0.26	0.38	0.39
	mBERT	Attn (1st)	0.4	0.49	0.58	0.6
		Attn (last)	0.4	0.49	0.58	0.6
		Flow	0.4	0.49	0.58	0.6
		Saliency	0.22	0.27	0.41	0.41
German	BERT	Attn (1st)	0.47	0.69	0.78	0.79
		Attn (last)	0.43	0.67	0.76	0.77
		Flow	0.46	0.69	0.78	0.79
		Saliency	0.17	0.39	0.5	0.5
	mBERT	Attn (1st)	0.46	0.72	0.78	0.8
		Attn (last)	0.46	0.72	0.78	0.8
		Flow	0.47	0.72	0.78	0.8
		Saliency	0.23	0.46	0.53	0.54
Russian	BERT	Attn (1st)	0.1	0.27	0.51	0.51
		Attn (last)	0.09	0.28	0.52	0.52
		Flow	0.09	0.26	0.51	0.51
		Saliency	0.08	0.24	0.5	0.5
	mBERT	Attn (1st)	0.1	0.28	0.51	0.51
		Attn (last)	0.11	0.28	0.51	0.51
		Flow	0.11	0.27	0.5	0.51
		Saliency	0.13	0.25	0.49	0.49