Native Language Prediction from Gaze: a Reproducibility Study

Lina Skerath
IT University of Copenhagen

Paulina Toborek
IT University of Copenhagen

Anita Zielińska
IT University of Copenhagen

Maria Barrett
IT University of Copenhagen
mbarrett@itu.dk

Rob van der Goot
IT University of Copenhagen
robv@itu.dk

Abstract

Numerous studies found that the linguistic properties of a person's native language affect the cognitive processing of other languages. However, only one study has shown that it was possible to identify the native language based on eye-tracking records of natural L2 reading using machine learning. A new corpus allows us to replicate these results on a more interrelated and larger set of native languages. Our results show that comparable classification performance is maintained despite using less data. However, analysis shows that the correlation between L2 eye movements and native language similarity may be more complex than the original study found.

1 Introduction

Research has shown that a speaker's native language can affect their learning and performance in a foreign language (Berkes and Flynn, 2012; Alonso, 2016; Cop et al., 2017). The eye movements of a reader, namely fixations and saccades, are a window to the online cognitive processing of text with milliseconds accurateness (Rayner, 1998). Native speakers of different languages may exhibit different eye movement patterns when reading a foreign language, with those reading in their native language making shorter and more frequent fixations while making longer fixations due to the increased cognitive load when reading in other languages (Hopp, 2010; Rayner et al., 2012; Berzak et al., 2022).

Several researchers have examined eye-movement patterns across different nationalities, exploring various aspects such as sentence reading times, fixation count, and saccade duration (Cop et al., 2015). Roberts and Siyanova-Chanturia (2013) showed that gaze data could be used for examining, e.g., reading processes, second language acquisition, and discourse processing, as well as give relevant insights into fields of second language acquisition and processing. Early research in Native Language Identification (Tsur and Rappoport, 2007) focused on the relationship between a person's native language and their writing in a second language, while Berzak et al. (2017) for the first time predicted a reader's native language using machine learning across four languages (Chinese, Japanese, Portuguese, and Spanish) using only eye-tracking features from natural reading in their second language (L2), English. The study leveraged the knowledge that different languages have unique features, such as word order, grammatical rules, and phonological features, that affect language processing in other languages.

Despite a general interest in eye-tracking corpora for L2 reading, e.g., (Cop et al., 2017), until recently, there has not been a publicly available dataset with enough languages to reproduce the results of Berzak et al. (2017). Berzak et al. (2017) used a subset of the licensed CELER dataset (Berzak et al., 2022) which is the largest eye-tracking corpus by the number of L2 readers encompassing five different native language backgrounds. The Multilingual Eye-movement COrpus (MECO) L2 dataset (Kuperman et al., 2022)[1] comprises English L2 reading by 12 different language backgrounds and allows replication of the findings by Berzak et al. (2017) on a different and larger set of languages which is why we employ the MECO dataset for this study.

In this study, we replicate the study by Berzak et al. (2017) and classify the native language of the reader from eye-tracking records of them reading English from another corpus.[2] We include readers from seven different language backgrounds that are more interrelated than the original study; the

[1] Publicly available at https://osf.io/q9h43/

[2] The code and data used in the project is publicly available at https://github.com/linaskerath/ANLP_project

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 152–159
July 10-12, 2023 ©2023 Association for Computational Linguistics

LANGUAGE	ISO	n PARTICIPANTS
Estonian	et	23
English	en	21
Finnish	fi	23
German	de	23
Hebrew	he	18
Italian	it	20
Spanish	es	21

Table 1: Number of participants by native language and language ISO code in the data set.

linguistic similarity of the languages used in this study is in the range of 0.64–0.89[3]. The original study did not explore languages in this range but only less similar languages (linguistic similarity <.5) plus one very similar language pair (linguistic similarity >.95).

2 Data

The MECO data was collected in 12 eye-tracking laboratories around the world. Participants were young adults ranging from 18 to 39 years old with high levels of L2 proficiency, which was ensured through English instruction in higher education. For more comprehensive information about the dataset, we refer to the authors' paper (Kuperman et al., 2022).

The MECO data set includes eye-tracking input gathered from native speakers of 12 languages recorded during reading an English encyclopedic text. Due to an insufficient number of participants in some of the cohorts, we used the subset of seven languages with the most participants. To avoid overfitting, we randomly undersampled 23 participants for the two largest cohorts, equivalent in size to the third largest group within the dataset as shown in Table 1. Berzak et al. (2017) used 36 to 37 readers for each language.

We only use the texts read by all the participants (also named "shared regime" in Berzak et al. (2017)). The total amount of words read per participant is 595 words, while the original study used 900 words. The feature set employed comprises three word-based measurements: First Fixation duration (FF), First Pass duration (FP) which is the sum of all fixations during the first pass reading of the word, and Total fixation duration (TF).

3 Methods

In this section, we describe the methods employed to replicate Berzak et al. (2017), giving a detailed description of the steps deviating from the setup of the original study.

3.1 Features

All data gaps encountered in the MECO dataset related to words marked as skipped by participants during reading, so it is legitimized to replace such shortages with zeros. Additionally, following the approach of the original research, we normalize all fixation times with the reading time of the entire sentence. The final data set consists of three fixation measures columns per word or cluster, where each row represents data collected from one person.

Words in Fixed Context (WFC) The WFC feature set considers the fixation times for specific words, and no aggregation is performed on the unigram level. The bigrams and trigrams fixation times are then obtained by simply summing values of unigrams that are a part of the interest area. Columns of the dataset consist of the 3 features for every n-gram in the corpus - 5364 features in total.

Syntactic Clusters (SC) In Berzak et al. (2017), syntactic features were obtained from the original Penn Treebank. As no manually annotated syntactic features are available for our data we use predicted syntactic information instead (described in detail in Appendix B). Following Berzak et al. (2017) we use the average FF, FP and TF over n-grams (n=1-3) of the UPOS labels, PTB POS tags, and UD dependency labels as features. For example: the average fixation time of a participant on the UPOS sequence ADV ADJ is a single feature.

Information Clusters (IC) Next to grouping the features by syntactic labels, the average fixation times were calculated for clusters created by the length of the words, measured as a number of characters. For bi- and trigrams, lengths of words were summed and thus clusters were created based on this sum.

3.2 Model

For interpretation, we compare to a majority class baseline. Following the original paper, we use a log-linear model to obtain the Native Language Identification from Reading (NLIR) performance as well as the model-based language similarity (3.3.2). We implement the model using scikit-learn (Pedregosa

[3] The calculation is explained in §3.3.1

	Shared regime		
Majority Class	15.44		
	unigrams	+bigrams	+trigrams
IC	47.52	48.19	48.86
SC	57.62	73.29	76.57
SC+IC	52.29	73.29	77.95
WFC	**81.29**	79.29	77.95

Table 2: NLIR results for log-linear model and majority class baseline.

et al., 2011) and use the 'lbfgs' solver in accordance with the original paper. A reader's native language encoded as a categorical variable is used as the model's target variable. We report our results based on 10-fold cross-validation. To preserve a similar distribution of languages in train and test data, we employ a stratified K-Folds split. We train the same model on the three feature sets described in the previous section and an additional combination of SC and IC feature sets.

To ensure comparability with the original paper despite the different amounts of languages, we analyze model performance with different amounts of languages. We train the model on each possible combination of languages and group them by the number of languages. We take the mean accuracy score of each group size and plot the results (figure 1). We note that our classes are slightly imbalanced, so arguably F1 could be a better metric but to compare to previous work and because the classes are almost balanced, we choose to use accuracy.

3.3 Similarity metrics

Berzak et al. (2014, 2017) suggest a link between English as a second language (ESL) production and linguistic similarities. To recreate the language similarity plots from the original study, we derive the same model-based metric and a cosine similarity based on syntactic and geographical features of a language.

3.3.1 Linguistic-based similarity

We use the same procedure and data as the original study to derive this similarity metric.The data is obtained from URIEL Typological Compendium (Littell et al., 2017a). Information selected is data derived from the World Atlas of Language Structures, features from Syntactic Structures of the World's Languages, and data from parsing the prose topological descriptions in Ethnologue. This information is supplemented by data on the languages belonging to different families, retrieved from Glottolog's world language tree. We use lang2vec (Littell et al., 2017b) for obtaining the complete feature vectors (with KNN completion). After truncating features with the same values among all languages,[4] we get a total of 189 features. The similarity scores between languages are then calculated as a cosine similarity of their feature vectors.

3.3.2 Model-based similarity

The model-based similarity captures native language similarities paralleled in reading patterns. In the same way as Berzak et al. (2017), we define "the classification uncertainty for a pair of native languages y and $y^{'}$ in our data collection D, as the average probability assigned by the NLIR classifier to one language given the other being the true native language." It is called English Reading Similarity (ERS) and is defined as:

$$ERS_{y,y'} = \frac{\sum\limits_{(x,y)\in D_y} p(y^{'}|x;\theta) + \sum\limits_{(x,y^{'})\in D_y^{'}} p(y|x;\theta)}{|D_y| + |D_y^{'}|}$$

The model, trained on all seven languages to perform NLIR, is used to extract language similarity. We separately feed test data sets for a single language y at a time and extract prediction probabilities for each other language $y^{'}$. Then a mean of the two language probabilities is calculated.

It is suggested that a higher classification uncertainty indicates greater language similarity. In figure 2 we plot the similarity metrics against each other to test this in the original study implied link.

4 Results

Table 2 presents the results for the baseline and the log-linear model when using 10-fold cross-validation. The model is trained and evaluated on all seven languages.

All variants of the model perform substantially better than the majority class baseline. Similarly to the results by Berzak et al. (2017), the model trained on the WFC feature set achieves the highest cross-validation accuracy (81.29%). While the model trained on syntactic and information cluster features improves with additional bi- and tri-grams, the words in the fixed context feature set do not follow this trend which differs from the original paper's results.

[4] Note that this can be considered non-standard, as the features of a language might impact the similarity between two other languages. We mainly used this strategy to follow the previous setup

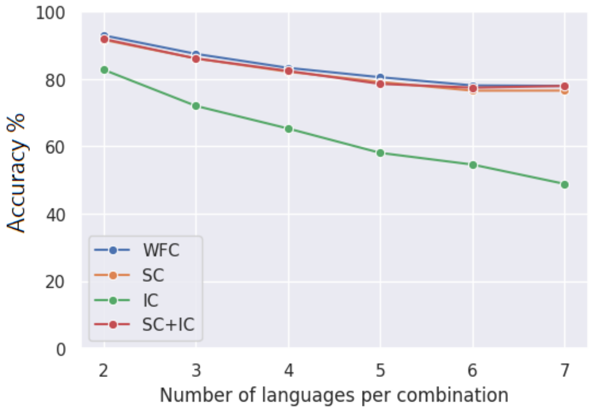


Figure 1: Mean performance of all combinations of languages using uni+bi+trigram features.

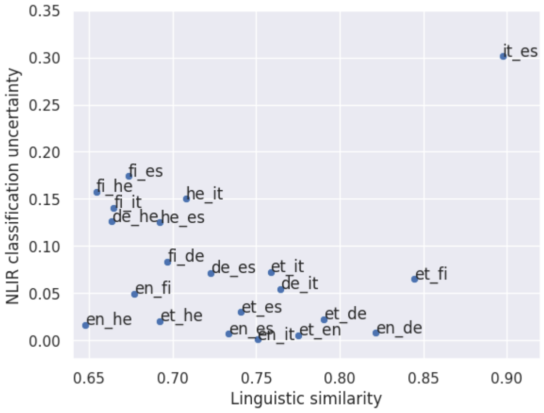
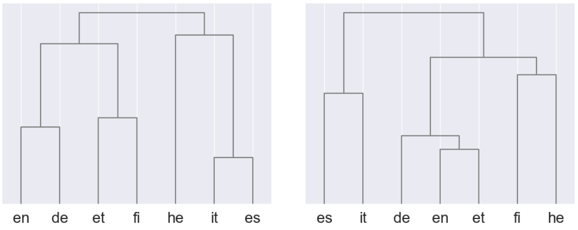


Figure 2: Linguistic similarities from URIEL against mean NLIR classification uncertainty of the unigram SC+IC model.



(a) Linguistic similarities

(b) Class uncertainty

Figure 3: Ward hierarchical clustering. Based on the unigrams SC+IC model.

Since the original study was done with a different number of languages, we investigate how the performance changes depending on the number of target classes. Figure 1 shows the changes in model performance depending on how many target classes it has. E.g., 3 on the x axis corresponds to a group of all combinations (C_7^3) of any three languages in the train set. The y-axis shows the mean performance of all classifiers in that group. The results of each classifier in a group vary, thus, we plot the mean performance. As expected, we see that for all feature sets the performance drops when the number of language increase.

5 Discussion

As evident from Table 2, our model seems to perform similarly to the original paper's results (Table 3, Appendix A). We can not compare these results directly due to the difference in languages, yet, for all combinations of four languages in our data set, we observe in Figure 1) that the average performance is 81 % (compared to 71% in Berzak et al. (2017). However, since we train our model with 3 more languages than the original study and still get similar results, we can confirm that machine learning models can pick up the differences in reading patterns of different native language readers. Contrary to the original paper, we do not see large improvements in performance with additional bigram and trigram features.

We also explore language similarity by looking at the suggested positive correlation between classification uncertainty and linguistic similarities. Results from Berzak et al. (2017) are included in Figure 5, Appendix A for convenience. The plot reproduced in Figure 2 does not seem to confirm this hypothesis as no clear trend is visible. We observe that the uncertainty when classifying native speakers vs. L2 reading is substantially lower (mean 0.01) than when distinguishing two groups of L2 readers from those of different native languages (mean 0.11). We also compute a correlation coefficient of 0.06 which does not indicate a significant correlation found by Berzak et al. Similarly, Ward hierarchical clustering for linguistic similarities and classification uncertainty, presented in Figure 3, does not present a closeness between grouping using either of these metrics. The plots have little overlaps on the set of languages we used, contrary to the original finding, see Figure 4, and share a little similarity both in terms of languages in each cluster and the general shape of the tree. This suggests that the relation between the English reading patterns and language similarities of the native language found by Berzak et al. (2017) may be more nuanced than the original plot (Figure 4, Appendix A) initially suggests.

6 Conclusion

We replicate the finding of Berzak et al. (2017) and are the first to confirm their finding that a reader's native language can be predicted from gaze patterns when reading English text. Having a larger set of more interrelated languages than the original study, we achieve comparable classification results supporting the suggested cross-linguistic influence from the native language to L2. Despite the satisfactory performance of the NLIR model, the results of investigating the relationship between reading patterns and linguistic similarity are not as straightforward. We believe the relation to be more nuanced than suggested as we are not able to replicate the same outcomes.

Acknowledgements

Maria Barrett is supported by a research grant (34437) from VILLUM FONDEN.

References

R.A. Alonso. 2016. *Crosslinguistic Influence in Second Language Acquisition*. G - Reference,Information and Interdisciplinary Subjects Series. Multilingual Matters.

Eva Berkes and Suzanne Flynn. 2012. Multilingualism: New perspectives on syntactic development.

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.

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. Celer: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*, 6:41–50.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 21–29, Ann Arbor, Michigan. 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*, 49:602–615.

Uschi Cop, Denis Drieghe, and Wouter Duyck. 2015. Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PLOS ONE*, 10(8):1–38.

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.

Holger Hopp. 2010. Ultimate attainment in l2 inflection: Performance similarities between non-native and native speakers. *Lingua*, 120:901–931.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acarturk, 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, Kaidi Lõo, Marco Marelli, and Kerem Usal. 2022. Text reading in english as a second language: Evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*, pages 1–35.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017a. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017b. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading*. Psychology Press.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Leah Roberts and Anna Siyanova-Chanturia. 2013. Using eye-tracking to investigate topics in l2 acquisition and l2 processing. *Studies in Second Language Acquisition*, 35.

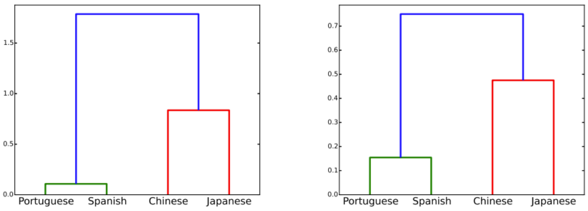
Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16, Prague, Czech Republic. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

A Results by Berzak et al. (2017)

	Shared regime		
Majority Class	25.52		
Random Clusters	22.76		
	unigrams	+bigrams	+trigrams
Information Clusters (IC)	41.38	44.14	46.21
Syntactic Clusters (SC)	45.52	57.24	58.62
Information Clusters (IC)	51.72	57.24	60.0
Words in Fixed Context (WFC)	64.14	68.28	**71.03**

Table 3: Native Language Identification from Reading results by Berzak et al. (2017)



(b) Linguistic tree

(c) English reading tree

Figure 4: Ward hierarchical clustering of linguistic similarities between languages and NLIR average pairwise classification uncertainties by Berzak et al. (2017)

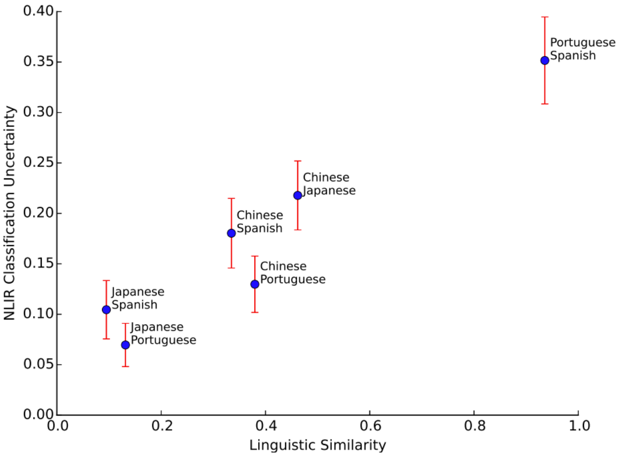


Figure 5: Linguistic similarities against mean NLIR classification uncertainty from Berzak et al. (2017)

B Obtaining Syntactic Annotations

We trained a multi-task MaChAmp model (van der Goot et al., 2021), including UPOS, PTB POS, lemmatization, morphological tagging, and dependency parsing. We used MaChAmp v0.4 with default settings, trained on the English Web Treebank v2.11 (because it has PTB tags and is English). It uses the combined (summed cross-entropy) loss of all tasks. We do not use the morphological tags and lemmas but include them for future work. All default hyperparameters are used and the default dev-split is used for model picking. We first ran the parser on the untokenized input but noticed that it quite commonly outputs the PUNCT label and corresponding relations to (end-of-sentence) words that have punctuation attached. So we pre-split using the BasicTokenizer from huggingface (which only separates punctuations) and use the labels of the words for the combined string. We compared mBERT (Devlin et al., 2019) with XLM-R Large (Conneau et al., 2020) and MLUKE (Ri et al., 2022). We compared their outputs on the MECO dataset manually and found the best performance with the XLM-R Large model (although MLUKE gets higher accuracies on EWT-dev).

C Limitations

The MECO dataset (Kuperman et al., 2022) is recorded at different labs following the same strict protocol. Nevertheless, location and experimenter effects may be confounding factors for the NLIR task. The CELER data (Berzak et al., 2022), used by (Berzak et al., 2017), seems to all be recorded at the same lab. Since we confirm their hypothesis, we do not see this as a fatal flaw in our study. There is no other available dataset that would allow us to replicate their finding.