Interpretable Machine Learning for Societal Language Identification: Modeling English and German Influences on Portuguese Heritage Language

Soroosh Akef[1,2] **Detmar Meurers**[3,2] **Amália Mendes**[1] **Patrick Rebuschat**[4,2]
[1]Center of Linguistics of the University of Lisbon, Portugal
[2]LEAD Graduate School and Research Network, University of Tübingen, Germany
[3]Leibniz Institute für Wissensmedien (IWM), Germany
[4]Lancaster University, United Kingdom
sorooshakef@edu.ulisboa.pt d.meurers@iwm-tuebingen.de
mendes@edu.ulisboa.pt p.rebuschat@lancaster.ac.uk

Abstract

This study leverages interpretable machine learning to investigate how different societal languages (SLs) influence the written production of Portuguese heritage language (HL) learners. Using a corpus of learner texts from adolescents in Germany and the UK, we systematically control for topic and proficiency level to isolate the cross-linguistic effects that each SL may exert on the HL. We automatically extract a wide range of linguistic complexity measures, including lexical, morphological, syntactic, discursive, and grammatical measures, and apply clustering-based undersampling to ensure balanced and representative data. Utilizing an explainable boosting machine, a class of inherently interpretable machine learning models, our approach identifies predictive patterns that discriminate between English- and German-influenced HL texts. The findings highlight distinct lexical and morphosyntactic patterns associated with each SL, with some patterns in the HL mirroring the structures of the SL. These results support the role of the SL in characterizing HL output. Beyond offering empirical evidence of cross-linguistic influence, this work demonstrates how interpretable machine learning can serve as an empirical test bed for language acquisition research.

1 Introduction

Cross-linguistic influence (CLI), or language transfer, broadly refers to the ways in which the linguistic representations of multilingual speakers interact with and affect one another. In the past several decades, this phenomenon has been a central issue of second language acquisition (SLA) research, as how a learner's L1 can shape the trajectory of L2 development has been extensively investigated (Odlin, 2022). Although the initial focus of CLI research was on the L1 transfer effect on L2, it is now believed that linguistic representations within the mind of a multilingual resemble a web, with complex interactions between all their linguistic systems (Macwhinney, 1987; McManus, 2021). The insights gained from this line of research have not only contributed to our understanding of the processes involved in language acquisition, but have also had implications for instructed SLA (McManus, 2019). Nevertheless, the focus of CLI research thus far has disproportionately been on L2 and its interaction with L1, with far less attention being given to CLI in other bilingual settings, particularly that of heritage language (HL) learners.

HL learners are individuals who grow up in an environment where a minority language is spoken at home while a dominant societal language (SL) is spoken in the broader community, possibly as a result of immigration (Benmamoun et al., 2013). Such learners often acquire their HL in naturalistic family settings during childhood, even as their formal education and daily social interactions are primarily conducted in the SL. Over time, the HL may develop differently than it would in a majority-language environment, resulting in a divergent outcome from that of native speakers who acquired their language in their home country (Bayram et al., 2019). While this divergence has been attributed to many factors, including the lower quality and quantity of input (Flores and Barbosa, 2014), the influence of the SL has been

Soroosh Akef, Detmar Meurers, Amália Mendes and Patrick Rebuschat. Interpretable Machine Learning for Societal Language Identification: Modeling English and German Influences on Portuguese Heritage Language. *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2025)*. University of Tartu Library. 50–62.

discussed as a contributor (Scontras et al., 2015) even though the empirical evidence for this influence has been mixed (van Osch, 2019; Torregrossa et al., 2023). This lack of conclusive evidence calls for more exploratory studies leveraging broad linguistic features that potentially capture the effect of the SL on HL production, particularly at various stages of development.

A critical gap involves understanding whether and how different SLs might variably influence the same HL. To date, comparisons that explicitly investigate how distinct SLs shape the development of a single HL have been limited, prompting Scontras et al. (2015) to call for such studies. Such comparisons could also contribute to our understanding of the impact of typological proximity, lexical overlap, and structural similarity on the development of the HL.

Consequently, our study aims to address this gap by examining the influence of two distinct, albeit typologically and genetically related, SLs (German and English) on the production of a single HL, European Portuguese. We employ a range of computational tools and methodologies, including the automatic extraction of linguistic complexity measures, topic modeling, clustering-based undersampling, and the application of interpretable machine learning models, to determine whether these models can reliably distinguish between HL texts produced in different SL contexts, a task we refer to as societal language identification (SLI). Specifically, we address the following research question: Can a machine learning model distinguish between texts produced by Portuguese HL learners with different SLs (English or German) using a wide range of linguistic complexity measures? By focusing on the types of linguistic complexity measures that distinguish between these two learner groups, we aim to identify patterns that may mirror tendencies of the respective SLs and in doing so, demonstrate the benefits of utilizing interpretable machine learning as an empirical test bed.

Beyond its theoretical implications, this task includes considerable practical significance: Insights gained from this line of research can inform the design of more personalized intelligent computer-assisted language learning (ICALL) systems that are geared to the unique needs of HL learners with different SL backgrounds. Drawing on the noticing hypothesis (Schmidt, 1990), which posits that mere exposure is insufficient for language acquisition and that learners must consciously pay attention to linguistic features of their input to acquire them, such systems can be designed to provide targeted exposure through input enhancement (Meurers et al., 2010) and input enrichment (Chinkina and Meurers, 2016), addressing specific areas of weakness. Taking different language backgrounds into account is also needed to draw valid inferences about learner competencies in Intelligent Tutoring Systems (Amaral and Meurers, 2008). Furthermore, understanding the specific ways in which SLs influence HL production is essential to ensure fairness in automatic language proficiency testing. Incorporating features that consistently predict proficiency across different SL or L1 backgrounds could mitigate potential biases that may unfairly disadvantage certain learner groups.

In the following sections, we begin with a review of the related work on CLI as it relates to HL acquisition and the task of native language identification (NLI), a task similar to, yet distinct from, the current task of SLI. Subsequently, we describe the methodology of our experiment, including the corpus of Portuguese HL texts, the automatic extraction of linguistic complexity measures, our attempt of controlling for text topic and balancing the data, and the interpretable machine learning approach utilized. This is followed by a presentation of the results, where we highlight the key findings related to the distinctions between English- and German-speaking HL learners. Finally, we discuss the theoretical and practical implications of these findings for the characterization of HLs, and we conclude with directions for future research.

2 Related Work

2.1 Cross-Linguistic Influence on the Heritage Language

CLI in bilingual development is a multifaceted phenomenon affecting HL acquisition across various linguistic domains. The lexicon is often considered to be a linguistic domain which is highly susceptible to CLI. In their investigation of English HL speakers with Hebrew as the SL, Gordon and Meir (2024) found no effect of CLI on morphosyntax, yet significant differences between the HL groups and the baseline group with regard to lexicon were observed. Specifically, her-

itage speakers exhibited minor lexical production errors influenced by Hebrew. Similarly, Böttcher and Zellers (2024) investigated how Russian HL speakers in contact with English or German increased their use of vocalic-nasal filler particles, a pattern reflecting the tendencies of the SL. Such effects are indicative of the subtle ways CLI can manifest itself in the HL.

On the other hand, while some research suggests that morphology and syntax are more resistant to CLI than the lexicon, other studies have found that the SL can influence HL morphosyntax: Meir and Janssen (2021) demonstrated how Russian HL speakers in contact with Dutch or Hebrew struggled to produce accusative and genitive morphology with the same accuracy as monolingual Russian speakers, concluding that differences in the mapping of functional features influence HL morphological acquisition. Cuza (2013) similarly demonstrated how the absence of subject-verb inversion in English influenced Spanish HL speakers, making them struggle with inversion in embedded questions. Likewise, Seo and Cuza (2024) found that Korean HL speakers in an English-dominant environment overused demonstratives and underused bare nouns, patterns mirroring English nominal structures. Furthermore, Brehmer and Usanova (2015) reported that Russian HL speakers in Germany exhibited increased verb-final structures, possibly as a result of German CLI, yet they preserved other HL-specific pragmatic patterns.

Meanwhile, Fridman et al. (2024) found CLI to be a main mechanism behind HL grammar maintenance in adults across multiple morphosyntactic phenomena (i.e., adjective–noun agreement, accusative case morphology, and numerical phrases) among Russian HL speakers in Hebrew and English environments. Notably, while CLI was found to be a major predictor of HL grammar maintenance, increased input and proficiency were found to modulate its effects. By contrast, other studies, such as Verkhovtceva et al. (2023), reported no clear evidence of CLI in HL morphosyntax, attributing the observed variation primarily to the age of onset of bilingualism. Similarly, Torregrossa et al. (2023) did not find CLI to significantly affect the performance of Portuguese HL children from three different SLs on a cloze-test targeting various linguistic structures, pointing to the variability in whether and how CLI manifests.

Despite the fact that many studies that attempt to isolate the role of CLI utilize monolingual speakers as the baseline group, Rothman et al. (2023) have criticized this approach due to its assumption that the HL is deficient and that its speakers must strive to conform to the monolingual norm. In lieu of this approach, one of the alternative approaches they have recommended is comparing bilingual groups from different SLs, which can allow us to capture possible differences in their language use as a result of CLI without the implication of HL deficiency.

2.2 Text-Based Native Language Identification

While the present study deals with identifying the SL in the context of HL development, text-based NLI is a closely related task, whose techniques can be transferred to SLI. NLI seeks to determine an author's L1 based on their productions in a specific L2. Although the goal in our task differs, both NLI and SLI involve identifying subtle linguistic fingerprints of previously acquired or concurrently acquiring languages on a target language.

NLI has become an established task in computational linguistics, as evidenced by several shared tasks in the last decade (Tetreault et al., 2013; Malmasi et al., 2017; M et al., 2018). Studies in NLI, surveyed comprehensively by Goswami et al. (2024), have explored a variety of feature sets and modeling approaches.

The features used for this task range from shallow features such as n-grams (Mohammadi et al., 2017) and part-of-speech (POS) information (Malmasi and Dras, 2018) to lexical features (Malmasi and Dras, 2014) and syntactic features (Bykh and Meurers, 2014). Nevertheless, Goswami et al. (2024) warn that despite the success of n-grams in NLI, their success may be attributed to capturing thematic differences likely to be present in texts produced by learners coming from different countries, as learners tend to make references to aspects of their home country in their texts, which an n-gram model can exploit. In essence, n-grams fail to utilize features that are informative about the language development of learners. This highlights the importance of using features which not only result in the best model performance, but which also validly characterize the construct being modeled, which can ultimately result in better model generalizability (Akef et al., 2024).

The models having been used for NLI resemble most other machine learning tasks, covering a range of traditional machine learning classifiers, such as SVMs (Bykh et al., 2013), logistic regression (Vajjala and Banerjee, 2017), and ensemble classifiers (Malmasi and Dras, 2018); deep learning approaches, such as gated recurrent unit (Bhargava et al., 2017), and long short-term memory (Mundotiya et al., 2018); as well as more recent approaches leveraging large language models (Zhang and Salle, 2023).

While the focus of the vast majority of NLI attempts has been on achieving superior accuracy, there exists tangible value in investing more effort in investigating whether the manner in which a given model makes its predictions aligns with theories conceptualizing the construct being modeled. Moreover, interpretable machine learning approaches, defined as algorithms that not only identify patterns in the data to perform a particular task but that can be studied to gain insights into and extract knowledge from the data (in contrast to end-to-end black-box models) (Murdoch et al., 2019), can serve as an empirical test bed to map the possible effects of a broad range of predictors in a way that is unfeasible using traditional statistical analysis techniques.

3 Methodology

3.1 Data

The data analyzed in this study originate from texts produced by HL learners as part of the annual EPE certificate examination[1], organized by the Camões Institute, which is administered to Portuguese HL learners residing abroad. The Camões Institute, an institution affiliated with the Portuguese Ministry of Foreign Affairs, is charged with promoting Portuguese language and culture worldwide. Through its educational programs, including community schools and language courses, the Institute supports Portuguese families abroad and ensures that their children maintain a connection to their linguistic heritage. This particular examination targets adolescents (aged 15–18) who have grown up in Germany or the UK and are receiving formal instruction in Portuguese as a HL.

Table 1 illustrates the distribution of the corpus across the two societal languages (German and English) and the three Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) proficiency levels of B1, B2, and C1. The corpus, containing a total of 472 texts with an average word count of 162.03, has a relatively balanced distribution across the CEFR levels for the German group while there are relatively fewer texts in the B2 and C1 levels in the English group.

To ensure that differences attributed to the SL are not merely as a result of possible topic difference, it was necessary to control for text topic prior to training. To this end, topic modeling using latent Dirichlet allocation (LDA) (Blei et al., 2003) was performed on the entire corpus using the Gensim Python library (Řehůřek and Sojka, 2010). By iteratively calculating the semantic coherence score (Mimno et al., 2011) of up to ten topics, the following nine topics were identified in the corpus based on the most representative words for each topic:

1. Personal life and relationships
2. Technologies, libraries, and youth
3. Travel and accommodation
4. Art, tourism, and cultural activities
5. Future and virtual reality
6. Books, culture, and leisure
7. Nature and outdoor photography
8. Tablets, education, and everyday tech
9. Work and projects

Subsequently, topics 1, 3, 4, and 6 were deemed similar enough to be grouped under one general topic of *Personal, cultural, and recreational life* to minimize data loss while adequately controlling for text topic. Subsequent to this step, a total of 298 texts belonging to this general topic were kept in the dataset, whose distribution across CEFR levels and SL is displayed in Table 2

By focusing on a single, thematically homogeneous subset of texts, we ensure that differences in linguistic complexity and structure are not confounded by text topic.

[1] https://www.instituto-camoes.pt/en/index.php?Itemid=2924

	B1	B2	C1	Total
English	90 (53.3%)	37 (21.9%)	42 (24.8%)	169 (100%)
German	102 (33.7%)	100 (33.0%)	101 (33.3%)	303 (100%)

Table 1: Distribution of texts by SL and CEFR proficiency level.

	B1	B2	C1	Total
English	76 (59.8%)	34 (26.8%)	17 (13.4%)	127 (100%)
German	46 (26.9%)	58 (33.9%)	67 (39.2%)	171 (100%)

Table 2: Distribution of texts on the selected general topic by SL and CEFR proficiency level.

3.2 Features

A total of 653 linguistic complexity features were automatically extracted from the texts partly using CTAP (Chen and Meurers, 2016; Weiss and Meurers, 2019), a web-based linguistic complexity analyzer which has been expanded to support a number of languages, including Portuguese (Demattos, 2020; Ribeiro-Flucht et al., 2024), and partly using custom annotators we developed to identify European Portuguese constructions using the rule-based matching of the spaCy Python library (Honnibal et al., 2020). Linguistic complexity is often defined in terms of the degree of variety and sophistication of a language instance (Wolfe-Quintero, 1998) or in terms of how challenging a language instance is (Ellis and Barkhuizen, 2005). However, the features used in this study vary in terms of the theoretical perspectives to complexity, including structural complexity measures, operationalized in terms of the number and variety of linguistic properties (Bulté and Housen, 2012; Pallotti, 2015) to measures of developmental complexity, such as age of acquisition, and processing complexity, such as concreteness. Table 3 demonstrates the distribution of these features across various classes, and the full list of features is available on the study's OSF repository[2].

Count-based features indicate the raw counts of various linguistic units, such as tokens, clauses, or particular syntactic structures. While these features could be categorized under syntactic complexity since longer linguistic units often imply higher syntactic complexity, the length-dependent nature of them necessitates different treatment from normalized syntactic features. Count-based features include measures such as the number of agent modifiers or the number of complex noun phrases.

Lexical features form the largest category of features in this study. They capture the sophistication and richness of the vocabulary by examining, for instance, various forms of type-token ratio (root, logarithmic, corrected, standard), as well as frequency-based measures such as word frequency per million. In addition to these traditional lexical features, psycholinguistic measures, such as age of acquisition and imageability, which stem from psycholinguistic experiments on how words are processed, are also included in this feature class.

On the other hand, syntactic features quantify aspects such as the frequency and depth of subordinate clauses, the presence of particular phrase types, or the mean length of clauses. For example, features including prepositional phrase types per token or the rate of subordination shed light on the learners' ability to produce more complex syntactic constructions.

Morphological features gauge the complexity resulting from inflectional and derivational processes. They provide information about how effectively learners manipulate the morphological structures of Portuguese, including person, number, tense, and mood markers. Examples include measures such as first person per word token or indicatives per word token.

Another class of complexity features extracted are discursive features, which measure cohesion at the text level. This class uses the frequency and variety of discourse markers as a feature characterizing the cohesiveness of the language.

Finally, this study utilizes a set of grammatical complexity features based on the occurrence of various European Portuguese constructions. Guided by the *Referencial Camões*[3], a benchmark specifying at which levels of proficiency specific

[2] https://osf.io/8gqud/

[3] https://www.instituto-camoes.pt/activity/centro-virtual/referencial-camoes-ple

Class	Count-based	Lexical	Syntactic	Discursive	Morphological	Grammatical
Count	182	241	73	42	32	83

Table 3: Count of features by class.

European Portuguese structures should be taught, these features are designed to serve as criterial features (Hawkins and Buttery, 2010) whose consistent use can be indicative of a learner having reached a specific proficiency level. While there could be overlap between this class and other linguistic complexity classes, they are classified separately due to their expected capacity to distinguish between proficiency levels.

By utilizing these diverse feature sets ranging from shallow token counts and POS categories to sophisticated lexical, morphological, syntactic, and language-specific grammatical complexity measures, we create a rich representation for each text, well suited to detecting differences in language use that may arise from the influence of the SL. Moreover, it aligns with our goal of employing an interpretable machine learning model, as we can better understand the ways in which the SL affects the HL across various linguistic domains.

3.2.1 Justifying Broad Linguistic Complexity Modeling

Criticism has been leveled against experiments such as the current study, in which a broad set of linguistic complexity measures extracted based on different theoretical frameworks are utilized to study linguistic phenomena, with Bulté et al. (2024) likening this approach to p-hacking. However, this critique mischaracterizes the intent and methodology of our approach, which is fundamentally data-driven and aims to discover patterns rather than simply confirm pre-existing theoretical assumptions. While we acknowledge the importance of careful selection of predictors for hypothesis testing, our methodology contributes to a different stage of the cycle of scientific progress, namely data-driven discovery and theory-informed interpretation.

To extend the analogy used by Jarvis (2010), where asserting the existence of CLI effects are likened to establishing the guilt of a defendent in a criminal trial, our approach is analogous to a detective investigating a crime. Rather than start with a single theory about the perpetrator's motive, the detective gathers all available evidence that might possibly offer a clue, from DNA samples and fingerprints to witness testimonies and purchase records. This broad data collection allows for the discovery of unexpected connections and the subsequent development of a more comprehensive understanding of the crime. Similarly, we cast a wide net in terms of linguistic features, drawing inspiration from various theoretical perspectives on what might be relevant to SL influence. Hence, we do not presuppose the primacy of any single theoretical framework, but rather allow the data itself, through machine learning, to reveal which features are most informative. Our approach, therefore, can be characterized as exploratory data analysis (EDA) for hypothesis generation rather than confirmatory analysis through hypothesis testing (Carmichael and Marron, 2018), both of which are essential steps of scientific progress.

The key difference between our approach and p-hacking lies in the purpose of feature selection. While p-hacking involves iteratively testing numerous hypotheses and selectively reporting only those that achieve statistical significance, our goal is not to confirm pre-defined hypotheses about specific features, but rather to explore the feature space and identify which linguistic features are most capable of characterizing CLI. Subsequently, this data-driven feature selection informs theoretical interpretation and model building, which has shown to result in better model accuracy and generalizability (Bykh and Meurers, 2016; Bykh et al., 2013; Akef et al., 2024).

3.3 Clustering-Based Downsampling

To ensure that both SLs were equally represented at each proficiency level and to prevent model biases arising from imbalanced class distributions, a clustering-based downsampling technique was employed (Lin et al., 2017). The corpus on the selected topic initially contained a larger number of texts produced by German-speakers relative to English-speakers, particularly at the B2 and C1 levels. Without adjusting for these discrepancies, the resulting model could be influenced more strongly by the SL with greater representa-

tion, making it difficult to attribute observed linguistic patterns to the SL rather than sampling imbalance.

To this end, the dataset was divided into English and German subgroups for each CEFR proficiency level, with the goal of downsampling the larger subgroup to match the size of the smaller subgroup. To ensure that the selected samples from the majority subgroup remained representative of its overall distribution, a two-step dimensionality reduction and clustering process was employed. Specifically, the scikit-learn (Pedregosa et al., 2018) implementation of the principal component analysis (PCA) algorithm was utilized to reduce the dimensionality of the feature space from 653 features to 10 principal components. PCA serves to capture the most significant variance in the data while mitigating the noise and potential curse of dimensionality that could adversely affect the downstream clustering step.

Following dimensionality reduction, K-means clustering (also from scikit-learn) was applied to the lower-dimensional data. The number of clusters (K) for K-means was set to the size of the minority subgroup: 46, 34, and 17 for levels B1, B2, and C1 respectively. By calculating pairwise distances between the texts and the cluster centroids, the sample with the smallest distance to each centroid was selected as its representative.

Finally, these selected samples from the majority group were combined with all samples from the minority group to form a balanced subset at each proficiency level. By repeating this procedure for each level and concatenating the balanced subsets, a new dataset was obtained in which English and German texts are equally represented at each proficiency level, as demonstrated in Table 4.

While other methods such as upsampling could also address class imbalance, downsampling was chosen here to preserve the variance in the data. Upsampling through simple duplication or synthetic generation of minority-class texts could introduce biased patterns and potentially result in unrepresentative interpretation of the model's use of features to distinguish between the two SLs.

3.4 Training

To model the influence of the SL on the HL, this study employs explainable boosting machines (EBMs) (Nori et al., 2019), a class of inherently interpretable machine learning models. EBMs are a type of generalized additive model (GAM) that leverage gradient boosting while maintaining a transparent structure. Consequently, EBMs construct predictions as a sum of shape functions for each individual feature and specified feature interactions. This architecture makes it possible to reliably identify which features and interactions play a more important role in the model's predictions, both globally and locally.

EBMs have been successfully applied in various domains, such as healthcare and finance, where model transparency and trustworthiness are paramount (Chen et al., 2023; Consiglio, 2023). Their ability to combine state-of-the-art predictive performance with interpretability has made them appealing for high-stakes decision-making. In the context of language learning research, EBMs offer the opportunity to gain insights from the data which would not be possible using deep learning or large ensemble methods due to their complex decision-making processes. By contrast, EBMs facilitate the attribution of model decisions to specific linguistic features.

In this study, the balanced dataset obtained after clustering-based downsampling served as the data for our EBM training. In the preprocessing stage, the variable *Proficiency* was specified as an ordinal categorical feature while the linguistic complexity measures were treated as continuous. As interactions between proficiency and other complexity features may reveal developmental patterns influenced by the SL, a set of pairwise interactions involving *Proficiency* and each complexity feature was explicitly specified. These interactions allowed the EBM to capture how the relationship between linguistic features and the SL differs across proficiency levels. Additionally, as neither complexity measures nor proficiency can validly characterize CLI on their own, all main effects (i.e., standalone complexity measures) were excluded in favor of limiting feature space dimensionality.

Model training was performed using a 5-fold stratified cross-validation procedure through scikit-learn. Each fold involves splitting the data into training and test subsets, training an EBM, as implemented in the InterpretML Python library (Nori et al., 2019) on the training set, and evaluating predictions on the test set. Following cross-validation, overall performance is calculated, and additional analyses are performed to examine performance by proficiency level.

	B1	B2	C1	Total
English	46 (47.4%)	34 (35.1%)	17 (17.5%)	97 (100%)
German	46 (47.4%)	34 (35.1%)	17 (17.5%)	97 (100%)

Table 4: Distribution of texts by SL and CEFR proficiency level after performing clustering-based downsampling.

After confirming the model's stability and predictive power using 5-fold cross-validation, the EBM was retrained on the entire balanced dataset. This final model facilitated the extraction of global feature importance measures. By interpreting these outputs, we were able to identify which complexity features at which proficiency levels best discriminate between HL texts produced by learners from different SLs.

4 Results and Discussion

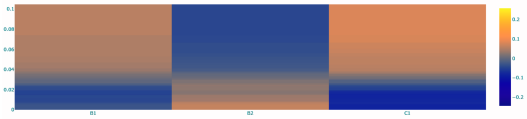
The EBM trained on the balanced subset of texts achieved a mean accuracy of 0.77 (±0.08) and a mean F1 score of 0.78 (±0.08) in 5-fold cross-validation, substantially above the random guess baseline of 0.5. Additionally, the model achieved a precision score of 0.76 (±0.06) and a recall score of 0.80 (±0.12). These performance metrics lend support to the SL influencing the characterization of HL output. Furthermore, analyzing the performance of the model at each proficiency level revealed that the best performance was achieved at level C1 (Table 5), indicating that SL-driven divergences in complexity features become more pronounced as learners' HL proficiency develops, possibly as a result of formal education in the SL.

	Accuracy	Precision	Recall	F1
B1	0.77	0.76	0.80	0.78
B2	0.75	0.73	0.79	0.76
C1	0.82	0.82	0.82	0.82

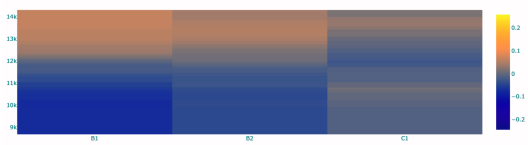
Table 5: Model performance by proficiency level based on out-of-fold predictions.

Extracting the most important features for EBM's distinction between the two SLs revealed potential traces of CLI across different linguistic domains (Table 6). However, to determine whether a group of linguistic features on average contributed more to the performance of the model, average feature importance for each class of features was calculated (Table 7), which revealed the greater role of morphological and lexical features, compared to the other classes.

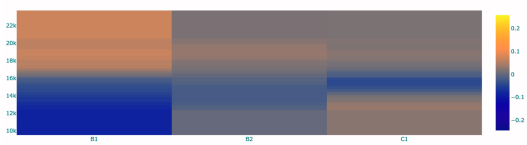
To zoom in on how these two groups of fea-



(a) Portuguese Vocabulary Profile (B1)



(b) SUBTLEX Word Frequency (SD AW)



(c) SUBTLEX Word Frequency (FW)

(d) SUBTLEX Contextual Diversity (FW)

Figure 1: Partial plots for the top four lexical features. Darker shades indicate a higher predicted likelihood of German as SL while lighter shades indicate a higher predicted likelihood of English as the SL. The figures have been post-processed for colorblind-friendliness; the original images are available in the study's OSF repository.

Feature	Category
Irregular Verbs in Imperfect Indicative (per verb token)	Grammatical
SUBTLEX Contextual Diversity (FW Token)	Lexical
Portuguese Vocabulary Profile (B1)	Lexical
Regular Verbs in Simple Past Indicative (per verb token)	Grammatical
Imperfect Tense (per verb token)	Morphological
SUBTLEX Word Frequency (SD AW Token)	Lexical
Number of Irregular Verbs in Imperfect Indicative	Count-based
SUBTLEX Word Frequency (FW Token)	Lexical
Infinitive Nominal Subordinate Clauses with Optative and Volitive Verbs	Count-based
SUBTLEX Word Frequency (SD FW Token)	Lexical
Difficult Connectives (per token)	Discursive
First Person (per word token)	Morphological
SUBTLEX Logarithmic Contextual Diversity (FW Token)	Lexical
Number of Agent Modifiers	Count-based
SD of Global Noun Overlap (lemma-based)	Discursive
Regular Verbs in Imperfect Indicative (per verb token)	Grammatical
Punctuation Density	Syntactic
Passive Verbs (per verb token)	Morphological
SUBTLEX Frequency Top 5000	Lexical
SUBTLEX Frequency Band 4	Lexical

Table 6: Top 20 most important features for the EBM model.

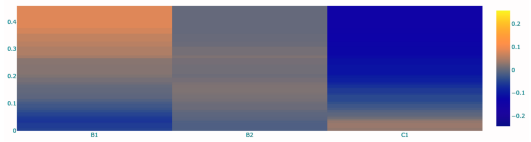
Class	Count-based	Lexical	Syntactic	Discursive	Morphological	Grammatical
Importance	13.47%	20.34%	15.64%	18.87%	21.32%	10.37%

Table 7: Average feature class importance.

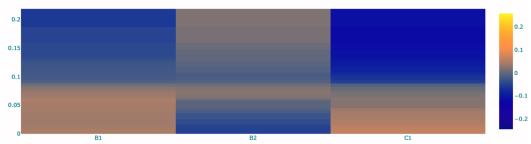
tures can capture possible CLI in texts produced by English- and German-speaking HL Portuguese learners, we took advantage of the additive structure of EBMs to visualize how specific features contribute to the prediction of the model (Figure 1). While distinct patterns in the top lexical complexity features for each SL are visible across proficiency classes, these differences seem to wane and become less pronounced as learners become more proficient, particularly visible in Figures 1b and 1d. This phenomenon could be indicative of the regularizing effect of higher proficiency on lexical choice in HL learners of different SLs. This assertion is consistent with English-speaking HL learners of Portuguese possibly leveraging cognates of the two languages in the earlier stages of development resulting in higher standard deviation of word frequency for all words (Figure 1b), a measure of linguistic diversity. Similarly, the sudden surge at level B2 and the subsequent drop at level C1 of the use of words characteristic of L2 textbooks at level B1 (Torigoe, 2017) by German-speaking learners (Figure 1a) is suggestive of different developmental trajectories among HL learners with distinct SLs.

We also visualized the contribution of the top morphological features to the model's predictions at each proficiency level (Figure 2). Similar to the patterns observed in lexical complexity features, there are distinct morphological preferences that appear to align with the learner's SL. For instance, English-speaking HL learners consistently exhibit a higher tendency to employ the passive voice across all proficiency levels, with the distinction between the two groups of learners regarding this feature becoming more pronounced at the C1 level (Figure 2c). This pattern may be explained by the structural similarity of the passive voice in English and Portuguese, as opposed to German, making it more accessible to learners whose SL is English. In contrast, German-speaking HL learners show a clear preference for using the imperfect tense and the first person as they become more proficient (Figures 2a and 2b). The preference for the imperfect tense among German-speaking learners may stem from the presence of a comparable

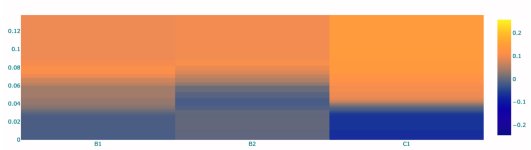
tense form in German, facilitating its transfer into Portuguese. The inclination toward first-person constructions by German-speaking learners could similarly be interpreted as consistent with their lack of preference for the passive voice. In contrast to lexical features which showed a tendency to converge as learners from distinct SLs become more proficient in their HL, the influence of morphosyntactic features follow the opposite trend, with learners' HL appearing to be influenced more heavily by the morphosyntactic properties of the SL at more advanced levels. An explanation for this could be that as learners progress through their HL classes, the SL, as their dominant language, also continues to become more entrenched as a result of formal education in the SL.



(a) Imperfect tense (per verb token)



(b) First Person (per word token)



(c) Passive verbs (per verb token)

Figure 2: Partial plots for the top three morphological features. Darker shades indicate a higher predicted likelihood of German as SL while lighter shades indicate a higher predicted likelihood of English as the SL. The figures have been post-processed for colorblind-friendliness; the original images are available in the study's OSF repository.

5 Conclusion

This study set out to explore CLI in HL learners of Portuguese by examining how the SL shapes patterns of lexical and morphosyntactic use, following the detection-based approach to CLI research (Jarvis, 2010). While our findings highlight certain trends, particularly with regard to lexical and morphological preferences, it is important to recognize that due to the exploratory nature of the study, these results offer only one perspective within a broader landscape of theoretical and empirical approaches. Rather than provide a definitive characterization of CLI in the context of HL, our aim was to explore how data-driven approaches, specifically interpretable machine learning, can be utilized to conduct scientific inquiry into CLI. Through more extensive datasets, more detailed typological comparisons, and closer engagement with CLI theory, subsequent investigations can refine our understanding of CLI, allowing us to move beyond preliminary evidence toward a richer, more comprehensive account of how the SL shapes the evolving linguistic knowledge of HL learners.

Acknowledgments

This work was developed within the scope of the project *Promoção da Aquisição e ensino do Português como Língua de Herança através de Ferramentas Digitais Inteligentes*, financed by the Foundation for Science and Technology - FCT of the Republic of Portugal and the Camões Institute. We would like to thank anonymous reviewers for their insightful comments on a previous version of this paper.

References

Soroosh Akef, Amália Mendes, Detmar Meurers, and Patrick Rebuschat. 2024. Investigating the generalizability of Portuguese readability assessment models trained using linguistic complexity features. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 332–341, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Luiz Amaral and Detmar Meurers. 2008. From recording linguistic competence to supporting inferences about language acquisition in context: Extending the conceptualization of student models for intelligent computer-assisted language learning. *Computer-Assisted Language Learning*, 21(4):323–338.

Fatih Bayram, Jason Rothman, Michael Iverson, Tanja Kupisch, David Miller, Eloi Puig-Mayenco, and Marit Westergaard. 2019. Differences in use without deficiencies in competence: passives in the Turkish and German of Turkish heritage speakers in Germany. *International Journal of Bilingual Education and Bilingualism*, 22(8):919–939. Publisher: Routledge _eprint: https://doi.org/10.1080/13670050.2017.1324403.

Elabbas Benmamoun, Silvina Montrul, and Maria Polinsky. 2013. Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical Linguistics*, 39(3-4):129–181.

Rupal Bhargava, Jaspreet Singh, Shivangi Arora, and Yashvardhan Sharma. 2017. Bits_pilani@inli-fire-2017: Indian native language identification using deep learning. In *FIRE*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Bernhard Brehmer and Irina Usanova. 2015. Lets fix it? In *Transfer Effects in Multilingual Language Development*, pages 161–188. John Benjamins.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2024. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*.

Bram Bulté and Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency*, Language Learning & Language Teaching, pages 21–46. John Benjamins Publishing Company, Amsterdam.

Serhiy Bykh and Detmar Meurers. 2014. Exploring syntactic features for native language identification: A variationist perspective on feature encoding and ensemble optimization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1962–1973, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Serhiy Bykh and Detmar Meurers. 2016. Advancing linguistic features and insights by label-informed feature grouping: An exploration in the context of native language identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 739–749, Osaka, Japan. The COLING 2016 Organizing Committee.

Serhiy Bykh, Sowmya Vajjala, Julia Krivanek, and Detmar Meurers. 2013. Combining shallow and linguistically motivated features in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 197–206, Atlanta, Georgia. Association for Computational Linguistics.

Marlene Böttcher and Margaret Zellers. 2024. Do you say uh or uhm? A cross-linguistic approach to filler particle use in heritage and majority speakers across three languages. *Frontiers in Psychology*, 15. Publisher: Frontiers.

Iain Carmichael and J. S. Marron. 2018. Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*, 1(1):117–138.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 113–119, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhi Chen, Sarah Tan, Urszula Chajewska, Cynthia Rudin, and Rich Caruana. 2023. Missing Values and Imputation in Healthcare Data: Can Interpretable Machine Learning Help? ArXiv:2304.11749 [cs] version: 1.

Maria Chinkina and Detmar Meurers. 2016. Linguistically aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 188–198, San Diego, CA. Association for Computational Linguistics.

Alessandro Consiglio. 2023. Model interpretability in credit insurance. Master's thesis, Instituto Superior de Economia e Gestão, March. Accepted: 2023-03-24T13:57:00Z.

Council of Europe. 2001. Common European Framework of References for Languages: Learning, teaching, assessment.

Alejandro Cuza. 2013. Crosslinguistic influence at the syntax proper: Interrogative subject–verb inversion in heritage Spanish. *International Journal of Bilingualism*, 17(1):71–96. Publisher: SAGE Publications Ltd.

Eric Demattos. 2020. Analyzing linguistic complexity of l2 portuguese for automatic proficiency classification. Master's thesis, Eberhard Karls University of Tübingen.

Rod Ellis and Gary Barkhuizen. 2005. *Analysing Learner Language*. Oxford University Press.

Cristina Flores and Pilar Barbosa. 2014. When reduced input leads to delayed acquisition: A study on the acquisition of clitic placement by portuguese heritage speakers. *International Journal of Bilingualism*, 18(3):304–325.

Clara Fridman, Maria Polinsky, and Natalia Meir. 2024. Cross-linguistic influence meets diminished input: A comparative study of heritage Russian in contact with Hebrew and English. *Second Language Research*, 40(3):675–708. Publisher: SAGE Publications Ltd.

Sidney Gordon and Natalia Meir. 2024. English as a heritage language: The effects of input patterns and contact with Hebrew. *International Journal of Bilingualism*, 28(3):353–373. Publisher: SAGE Publications Ltd.

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native language identification in texts: A survey. In

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3149–3160, Mexico City, Mexico. Association for Computational Linguistics.

John A. Hawkins and Paula Buttery. 2010. Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal*, 1:e5.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Scott Jarvis. 2010. Comparison-based and detection-based approaches to transfer research. *EUROSLA Yearbook*, 10(1):169–192.

Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. 2017. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409-410:17–26.

Anand Kumar M, Barathi Ganesh H. B., Ajay S. G, and Soman K. P. 2018. Overview of the second shared task on indian native language identification (INLI). In *Working Notes of FIRE 2018 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 6-9, 2018*, volume 2266 of *CEUR Workshop Proceedings*, pages 39–50. CEUR-WS.org.

Brian Macwhinney. 1987. *The Competition Model*, pages 249–308. Lawrence Erlbaum.

Shervin Malmasi and Mark Dras. 2014. Finnish native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, pages 139–144, Melbourne, Australia.

Shervin Malmasi and Mark Dras. 2018. Native language identification with classifier stacking and ensembles. *Computational Linguistics*, 44(3):403–446.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark. Association for Computational Linguistics.

Kevin McManus. 2019. Awareness of L1 form-meaning mappings can reduce crosslinguistic effects in L2 grammatical learning. *Language Awareness*, 28(2):114–138.

Kevin McManus. 2021. *Crosslinguistic Influence and Second Language Learning*. Routledge, New York.

Natalia Meir and Bibi Janssen. 2021. Child Heritage Language Development: An Interplay Between Cross-Linguistic Influence and Language-External Factors. *Frontiers in Psychology*, 12. Publisher: Frontiers.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Los Angeles, California. Association for Computational Linguistics.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Elham Mohammadi, Hadi Veisi, and Hessam Amini. 2017. Native language identification using a mixture of character and word n-grams. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 210–216, Copenhagen, Denmark. Association for Computational Linguistics.

Rajesh Kumar Mundotiya, Manish Singh, and Anil Kumar Singh. 2018. Nlprl@inli-2018: Hybrid gated lstm-cnn model for indian native language identification. In *FIRE*.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Terence Odlin. 2022. *Explorations of Language Transfer*. Multilingual Matters, Bristol, Blue Ridge Summit.

Brechje van Osch. 2019. Vulnerability and cross-linguistic influence in heritage spanish: Comparing different majority languages. *Heritage Language Journal*, 16(3):340 – 366.

Gabriele Pallotti. 2015. A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134. Publisher: SAGE Publications Ltd.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine learning in python.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In

Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta. ELRA.

Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2024. Explainable AI in language learning: Linking empirical evidence and theoretical concepts in proficiency and readability modeling of Portuguese. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 199–209, Mexico City, Mexico. Association for Computational Linguistics.

Jason Rothman, Fatih Bayram, Vincent DeLuca, Grazia Di Pisa, Jon Andoni Duñabeitia, Khadij Gharibi, Jiuzhou Hao, Nadine Kolb, Maki Kubota, Tanja Kupisch, Tim Laméris, Alicia Luque, Brechje van Osch, Sergio Miguel Pereira Soares, Yanina Prystauka, Deniz Tat, Aleksandra Tomić, Toms Voits, and Stefanie Wulff. 2023. Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Applied Psycholinguistics*, 44(3):316–329.

Richard W. Schmidt. 1990. The role of consciousness in second language learning1. *Applied Linguistics*, 11(2):129–158.

Gregory Scontras, Zuzanna Fuchs, and Maria Polinsky. 2015. Heritage language and linguistic theory. *Frontiers in Psychology*, 6. Publisher: Frontiers.

Yuhyeon Seo and Alejandro Cuza. 2024. On the production of bare nouns and case marking in Korean heritage speakers in contact with English. *Lingua*, 311:103826.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia. Association for Computational Linguistics.

Shintaro Torigoe. 2017. Portuguese Vocabulary Profile: uma lista de vocabulário a aprendentes do PL2/PLE, baseada nos corpora de aprendentes e de livros de ensino. *Revista da Associação Portuguesa de Linguística*, 3:387–400.

Jacopo Torregrossa, Cristina Flores, and Esther Rinke. 2023. What modulates the acquisition of difficult structures in a heritage language? a study on portuguese in contact with french, german and italian. *Bilingualism: Language and Cognition*, 26(1):179–192.

Sowmya Vajjala and Sagnik Banerjee. 2017. A study of n-gram and embedding representations for native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–248, Copenhagen, Denmark. Association for Computational Linguistics.

Tatiana Verkhovtceva, Maria Polinsky, and Natalia Meir. 2023. Cross-linguistic influence, limited input, or working-memory limitations: The morphosyntax of agreement and concord in Heritage Russian. *Applied Psycholinguistics*, 44(5):941–968.

Zarah Weiss and Detmar Meurers. 2019. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 380–393, Florence, Italy. Association for Computational Linguistics.

Kathryn Elizabeth Wolfe-Quintero. 1998. *Second language development in writing : measures of fluency, accuracy, & complexity*. University of Hawai'i, Second Language Teaching & Curriculum Center.

Wei Zhang and Alexandre Salle. 2023. Native Language Identification with Large Language Models. ArXiv:2312.07819 [cs].