Etymology Can Inform Lexical Complexity Prediction

Noah-Manuel Michael, Anna Hülsing, Andrea Horbach
Kiel University, Germany
Leibniz Institute for Science and Mathematics Education, Kiel, Germany
{michael|huelsing|horbach}@ipn.uni-kiel.de

Abstract

In Lexical Complexity Prediction (LCP) for second-language learners, cross-linguistic cognacy and borrowing relationships are commonly approximated using surface similarity. However, historical sound changes and orthographic differences can obscure etymological relationships that still facilitate vocabulary production. We investigate whether etymology-based features improve LCP for learners of English with German, Spanish, or Mandarin as source languages, using the 2026 BEA shared-task dataset. We expect etymological information to be most informative for German and Spanish, where historical relationships with English are frequent but often obscured by divergence. Mining Wiktionary etymologies, we derive nine etymology features for source–target word pairs, evaluated alongside surface-similarity and common lexical features in a stacked `RoBERTa-LightGBM` architecture. Across most configurations, adding etymology information significantly improves prediction performance for all three languages. For German and Spanish, the clearest contribution occurs for word pairs with normalized Levenshtein distances (NLDs) between 0.3 and 0.7, where surface similarity becomes less reliable, but common lexical features are not yet dominant. For Mandarin, etymology features remain useful, though less localized across NLD ranges. Our findings indicate that etymology effectively complements standard LCP approaches when surface similarity becomes less informative.

1 Introduction

Lexical Complexity Prediction (LCP) is a central task in educational NLP that aims to estimate how difficult individual words are for second-language (L2) learners (North et al., 2023). Accurate LCP is crucial for a range of downstream applications such as automated lexical simplification, readability assessment, and adaptive language-learning systems.

Lexical complexity is multidimensional and depends not only on properties of the target word, such as its frequency or length, but also on learner-related factors (Koutamanis et al., 2024). One such factor is the learner's first language (L1) through mechanisms such as cognate facilitation, where cognates are defined as words that share form and meaning across languages. Because of these similarities, they may be easier to acquire, retain, and retrieve (Otwinowska and Szewczyk, 2019).

Accordingly, cognacy has previously been approximated by surface similarity (Schepens et al., 2012; Mayer Martins et al., 2026). For example, German *Information* and Spanish *información* visibly resemble English *information*, whereas Mandarin 信息 *xìnxī* does not. However, following the historical-linguistic definition of cognacy, where cognates descend from a common origin (Crystal, 2008), surface similarity can only be an imperfect proxy, since regular sound changes, language-specific phonological adaptation, and differing orthographic conventions can obscure etymological relationships that may remain accessible to learners. For instance, English *father* and Spanish *padre*, as well as English *tongue* and German *Zunge*, are historically related despite their divergent surface forms. Similar considerations apply to lexical borrowings that have undergone phonological or orthographic adaptation in the respective languages.

We therefore argue that etymological information may provide an additional signal for computationally modeling lexical complexity that complements surface similarity. In particular, etymology-based features may be most informative for source–target word pairs with intermediate normalized Levenshtein distances (NLDs), for which the surface forms are neither sufficiently similar to provide a

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 173–196, November 19–20, 2026, University of Tartu Library

clear signal nor sufficiently dissimilar to rule out an etymological relationship. We expect this pattern to be clearest for German and Spanish, which offer comparatively many historically related forms due to their phylogenetic relatedness and centuries of shared culture and history facilitating language exchange. For Mandarin, a phylogenetically unrelated language to English, we expect weaker effects due to a comparatively smaller number of etymologically related words. If learners can exploit recurring cross-linguistic correspondences or knowledge of shared roots and borrowing histories, such relationships may facilitate the retention and subsequent production of English target words, resulting in lower experienced lexical complexity.

The 2026 BEA shared-task dataset on Vocabulary Difficulty Prediction for English learners is particularly suitable for our purpose because it measures the difficulty of producing an English target word from contextual information such as the source-language equivalent, the source-language word in context, and the first letter of the English target word plus a length indicator (e.g., a _ _ _ _ _ for *auburn*), rather than mere comprehension (Felice and Skidmore, 2026).[1] Michael et al. (2026) show that vocabulary items' productive difficulty exhibited by learners differs from receptive difficulty as intended by second language acquisition experts designing graded learning materials. The BEA setting allows us to investigate whether etymological relationships between L1 and English words affect learners' ability to retrieve and actively produce the target vocabulary.

To examine this question, we derive nine etymology-based features from Wiktionary entries for source–target word pairs, using `GPT-5.6-Luna` to parse the relevant etymological information. We investigate whether these features improve LCP performance per L1, whether they capture information distinct from surface similarity and other common lexical features, and whether their contribution differs across L1s. Specifically, we address the following research questions:

- RQ1: What is the predictive potential of etymology features in LCP, both alone and combined with existing features?
- RQ2: To what extent do etymology features contribute information distinct from features previously employed in LCP?
- RQ3: Is there a specific NLD range where etymology features are most effective, and does it vary across L1s?
- RQ4: To what extent do etymology features, alone or combined with existing features, further improve a transformer baseline in LCP?

Our results show that etymology features can inform LCP, particularly for German and Spanish source–target word pairs with intermediate NLDs (0.3–0.7). For Mandarin, etymology features remain helpful overall, but as expected for a more lexically distant L1, their contribution is less clearly localized across NLD ranges. Common lexical features are generally more competitive for this subset.

We make our code available at `https://github.com/noahmanu/etymology4lcp`.

2 Related Work

In the following, we first lay out the development of the L1-aware BEA Vocabulary Difficulty Prediction task. We then elaborate on the mechanism of cognate facilitation and the role of etymology information in LCP thus far.

Development of the BEA Vocabulary Difficulty Prediction Task. The 2026 BEA Shared Task is the cumulation of over a decade of research in automatically predicting word difficulty: Complex Word Identification is the binary task that decides which words in a given sentence are complex and therefore require simplification (Paetzold and Specia, 2016). The relevance of this task is underlined by it being the topic of several shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018; Ortiz-Zambrano and Montejo-Ráez, 2020). Instead of the binary Complex Word Identification label, Shardlow et al. (2020) proposed a continuous label, which allows annotators to make more fine-grained complexity decisions. This task became known as LCP. It was the sole focus of the SemEval 2021 Shared Task (Shardlow et al., 2021) and was later combined with lexical simplification in the BEA 2024 Multilingual Lexical Simplification Pipeline Shared Task (Shardlow et al., 2024). Vocabulary Difficulty Estimation is another related task, often framed as a CEFR-level classification task of words both in context (Bannò et al., 2025) and out of context (Alfter, 2024). The 2026 BEA Shared Task on Vocabulary Difficulty Prediction for English learners aims to automatically estimate

[1] In the context of this study, we treat lexical complexity and vocabulary difficulty synonymously.

a continuous difficulty target quantifying the likelihood of an average learner being able to produce target vocabulary given certain contextual information (Felice and Skidmore, 2026). This shared task distinguishes between learners of English with different L1s (Spanish, German, and Mandarin), building on prior work on L1-aware difficulty prediction (Palmero Aprosio et al., 2020; Skidmore et al., 2025).

The Role of L1 in LCP. While early Complex Word Identification and LCP work largely treated complexity as universal, later studies show that it differs between native and non-native readers (Gooding et al., 2021; North and Zampieri, 2023) and, crucially, between L2 learners with different L1 backgrounds (Ide et al., 2023). This motivates L1-aware prediction, as adopted by the 2026 BEA Shared Task.

Cognate Facilitation. Cognate facilitation effects are well known in psycholinguistic research on bilingualism and describe learners' general ability to recognize that two words from different languages sharing formal resemblance are "somehow the same thing" (Carroll, 1992), facilitating the comprehension, retention, and production of cross-linguistically similar items (Otwinowska and Szewczyk, 2019; Schepens et al., 2020; Woumans et al., 2021; Engemann and Radetzky, 2024). Accordingly, cognacy thus far has been estimated based on surface similarity. For instance, Beijering et al. (2008) and Schepens et al. (2012) used Levenshtein distance measures to model lexical distance and found robust correlations with bilingual speakers' distance judgments. More recently, Mayer Martins et al. (2026) participated in the 2026 BEA Shared Task and proposed a cosine similarity feature derived from TF-IDF-weighted n-gram vectors between word pairs as a proxy for cognacy. Löfflad and Kanli (2026) propose Levenshtein distance as a proxy for cognacy in L1-aware Automatic Readability Assessment.

Etymology in LCP. In historical linguistics, in contrast, a cognate is "a linguistic form which is historically derived from the same source as another [...] form" (Crystal, 2008). We argue that these historical developments may be accessible to learners to some degree, even if surface forms diverge, and may therefore facilitate LCP.

As Tack (2021) points out, "the link between perceived difficulty and the word's etymology has not yet been thoroughly investigated." So far, only tentative approaches at informing LCP with etymology information have been made. These include Tack's exploration of features based on Etymological WordNet, which, similarly to our approach, is mined from Wiktionary (de Melo, 2014), as well as the work of Philippov et al. (2026), who note the use of an etymology feature but provide no further details. To the best of our knowledge, we are the first to deliver an L1-aware, systematic evaluation of etymology features' impact on LCP performance.

3 Data

This section describes the dataset used in our experiments and the preprocessing we apply.

3.1 Dataset

We conduct our experiments on the 2026 BEA shared-task dataset on Vocabulary Difficulty Prediction for English learners (Felice and Skidmore, 2026), which is based on data collected in the context of the Knowledge-based Vocabulary Lists project (Schmitt et al., 2024). For that project, learners were asked to produce an English target word given its source-language equivalent(s) and contextual information including a source-language example sentence and the first letter of the English target word plus a length indicator (e.g., a _ _ _ _ _ for *auburn*). The dataset contains 7,516 parallel data points for each of the three source languages (German, Spanish, and Mandarin). Each source–target word pair is associated with a continuous Generalized Linear Mixed Model-derived lexical complexity score, reflecting the likelihood that a learner of average ability from the corresponding L1 population can produce the English target word given the contextual information. Higher scores indicate lower complexity.

3.2 Preprocessing

To maximize the number of instances available for our experiments, we pool the original training, development, and test splits across all source languages used in the shared task into a single dataset. This enables us to use a five-fold cross-validation setup (see Section 5.1 for more details).

We then perform the following preprocessing steps to prepare the data for feature computation: For the English target words, we only perform lowercasing. The L1 source-word column, on

the other hand, often contains multiple comma-separated alternatives as well as systematic additional annotations, such as argument placeholders or hints indicating which potential English target word is not intended. We therefore first apply a cleaning pipeline that removes bracketed annotations, hyphens, slashes, and language-specific grammatical markers while explicitly preserving comma-separated translation alternatives. Second, to create a fully standardized L1 representation, we then lowercase the text, convert Chinese characters to pinyin to enable the computation of surface-similarity features, and remove all diacritical marks (e.g., umlauts in German, accents in Spanish, and tone marks in Mandarin). Last, we identify the L1 alternative closest to the English target word based on their NLDs (scaled by the geometric mean of the word-pair lengths) and use this selected alternative to compute all downstream features.

4 Features

In the following sections, we introduce the three groups of manually engineered linguistic features whose contributions to LCP performance we evaluate alongside contextual predictions generated by `XLM-RoBERTa-base`:[2] **C**ommon lexical features, **S**imilarity features, and **E**tymology features based on etymological information extracted from Wiktionary. Table 1 provides an overview of the feature groups, which we describe in more detail below. Appendix B illustrates the possible values of each individual feature. We also briefly explain our reasoning for including an L1 feature.

4.1 Common Lexical Features

We define common lexical features as features that have repeatedly proven useful in previous LCP studies and that are not covered by our similarity feature group. We select word frequency, target word length, the estimated number of syllables in the target word, and its POS to serve as a representative subset of this family (North et al., 2023).

4.2 Similarity Features

While similarity features are also well established in LCP (Shardlow et al., 2021; Bestgen, 2026; Robaina et al., 2026), we treat them as a separate feature group because they constitute the conventional surface-form-based proxy for cross-linguistic relatedness against which we compare our etymology features. This distinction allows us to assess whether etymology captures information beyond that provided by orthographic and phonological similarity measures. As representatives of this feature group, we select raw and normalized Levenshtein distances (scaled either by maximum word length or the geometric mean of the word-pair lengths) computed on both orthographic forms and IPA transcriptions, Jaro and Jaro–Winkler distances, and cosine similarity across TF–IDF-weighted character n-gram vectors, as introduced by Mayer Martins et al. (2026).

4.3 Etymology Features

We introduce a novel family of etymology-based features designed to capture the historical, etymological, and cross-lingual connections between L1 source and English target words. We expect them to be informative for lexical complexity, particularly when regular sound changes, language-specific phonological adaptation, or differing orthographic conventions obscure etymological relationships at the surface level.

Generation of Etymology Features. We derive the etymology features from etymological information extracted from Wiktionary (Ylonen, 2022).[3] For each source–target word pair, we collect the relevant etymology sections from both the English edition and, where available, the edition corresponding to the respective source language. We then use `GPT-5.6-Luna` as a zero-shot feature annotator to parse this information and generate a `JSON` object containing the feature values. Brysbaert et al. (2025) and Philippov et al. (2026) have shown that using LLM-generated word familiarity estimates as features for LCP achieves improved performance. Therefore, where there is no etymology information available, we ask the LLM to fill in from its own knowledge. We provide the prompt we use in Appendix C and an overview of how many items have etymology information available in Appendix D. Coverage is comprehensive for English target words as well as German and Spanish source words. However, it is markedly lower for Mandarin, where etymological information is available for only approximately 10% of the source words. Following a manual inspection of 120 Mandarin instances,

[2] Corresponds to the baseline model used in the BEA Shared Task (Felice and Skidmore, 2026). We will henceforth shorten `XLM-RoBERTa-base` to `RoBERTa`.

[3] `https://kaikki.org`, last accessed 2026/10/30.

Family	Feature	Description
Common	log_frequency	Frequency-based lexical familiarity measure (SUBTLEX-UK)
	target_word_length	Number of characters in the English target word
	num_syllables	Estimated number of syllables in the target word
	pos_tag	Part-of-speech category of the target word
Similarity	char_edit_dist	Raw Levenshtein distance between source and target word
	char_edit_dist_norm	Maximum length-normalized Levenshtein distance between source and target word
	char_edit_dist_geo_norm	Geometric mean-normalized Lev. distance between source and target word (Mayer Martins et al., 2026)
	phon_edit_dist	Raw Levenshtein distance between IPA transcriptions of source and target word
	phon_edit_dist_norm	Maximum length-normalized Levenshtein distance between IPA transcriptions of source and target word
	phon_edit_dist_geo_norm	Geometric mean-normalized Levenshtein distance between IPA transcriptions of source and target word
	jaro_dist	Jaro distance between source and target word
	jaro_winkler_dist	Jaro-Winkler distance between source and target word
	soundex_dist	Raw Levenshtein distance between soundex codes of source and target words
	metaphone_dist_norm	Maximum length-normalized Levenshtein distance between metaphone codes of source and target words
	char_cosine_sim	Cosine similarity over TF-IDF-weighted character n-gram vectors (Mayer Martins et al., 2026)
Etymology	lca_language	Lowest common ancestral language stage (strictly excluding borrowing)
	same_proto_word	Whether both words descend from the exact same reconstructed proto-word
	same_proto_root	Whether both words share a broader root, but not the exact proto-word
	borrowed_from_source	Whether the English word was directly borrowed from the L1
	borrowed_from_target	Whether the L1 word was directly borrowed from English
	borrowed_from_third	Whether the L1 word and the English word were both borrowed from the same third language
	english_word_origin	Immediate taxonomic branch from which English acquired the word
	origin_matches_source_family	Whether the English word's immediate origin aligns with the L1 family
	transmission_path	Complexity of the word's transmission path into English

Table 1: Feature families, feature names, and descriptions. Feature value sets can be found in Appendix B.

we observe that the available etymological information often pertains to borrowed vocabulary, whereas native Chinese terminology is rarely covered. Because these borrowed terms align with our primary subject of interest, we maintain that our feature generation pipeline is sufficiently robust to produce accurate feature values for the Chinese subset.

Features Touching on Cognacy. To model the depth of the genetic lineage between the word pair, we extract the lowest common ancestor language, representing the youngest ancestral language stage at which the two words were still a single, indistinguishable lexical item. A closer shared ancestor, such as West Germanic compared to Proto-Indo-European, can imply higher degrees of structural and semantic similarity, thereby lowering the cognitive load for the learner. For instance, while English and German are both West Germanic languages, English *egg* and German *Ei* share a Proto-Germanic ancestor **ajją*, but not a Proto-West-Germanic one **aij*, because English lost its native West Germanic term *ey* and replaced it with the Old Norse borrowing *egg*.

We further refine this genetic linkage by determining whether the words descend from the exact same proto-word or merely share a foundational proto-root. Descending from a fully formed ancestral word acts as a strong cross-lingual anchor through true cognacy. For example, English *fish* and Spanish *pez* share a proto-word because both are inherited directly from the exact same Proto-Indo-European noun **péysks*. In contrast, English *youth* and Spanish *juventud* share only a proto-root. While both derive from the underlying Proto-Indo-European root for 'young' (**yeu-*), they do not descend from a single ancestral noun; rather, they were formed independently when the Germanic and Italic branches appended different abstract suffixes to that root. Capturing this distinction allows our models to differentiate between the maximum transparency of inheriting an entire word and the morphological cue of sharing ancestral building blocks.

Features Touching on Borrowing Relationships. To account for direct vocabulary transfer, we include two binary features indicating whether the English word was borrowed from the source language (e.g., *doppelganger* for a German-speaking learner) or whether the source word was borrowed from English (e.g., 沙拉 *shālā* 'salad' for a Mandarin-speaking learner). We also include a third feature indicating whether both the source and target words were borrowed from the same third language (e.g., English *tile* and German *Ziegel*, both derived from Latin *tēgula*).

Furthermore, we classify the taxonomic origin of the English word (i.e., Germanic, Romance, Sinitic, or Other) to capture the immediate genealogical branch from which an English word originates. In particular, Romance-derived vocabulary often corresponds to more formal or academic registers and may therefore be more complex than native Germanic vocabulary, irrespective of the learner's L1. We complement this feature with a binary indica-

tor of whether the English word's origin matches the source language family. This feature captures the possibility that words derived from a branch related to the learner's L1 may exhibit more familiar morphological patterns or regular correspondences. Thus, when German- or Spanish-speaking learners encounter Germanic- or Romance-derived English words, respectively, this familiarity may reduce the cognitive load involved in producing the word, even in the absence of direct cognacy.

Finally, we categorize the transmission path through which the word entered the English language (i.e., inheritance, direct borrowing, or indirect borrowing through one or more intermediary languages). Words that passed through multiple intermediary languages often undergo substantial semantic or orthographic shifts, obscuring their origins and potentially increasing their difficulty relative to natively inherited or directly borrowed vocabulary.

4.4 Inclusion of an L1 Feature

As the complexity-score distributions in the BEA shared-task dataset were derived from different population samples for each L1 (Skidmore et al., 2025), score values are not necessarily directly comparable across source languages. However, training on data pooled across L1s has generally been found to improve performance (Skidmore et al., 2025; Felice and Skidmore, 2026); we therefore adopt this approach. To enable our single multilingual model to account for language-specific differences in the interpretation and distribution of complexity scores, we also include L1 as an input feature in most configurations.

5 Method

Here, we describe our model architecture and the evaluation metrics we use. We also briefly introduce our experimental setup and our analyses.

5.1 Model Architecture

We conduct two complementary sets of experiments. First, to assess the predictive potential of etymology features in isolation and in combination with other features previously used in LCP, we train and evaluate `LightGBM` models using different feature group configurations. We choose `LightGBM` because tree-based models have consistently proven effective for LCP (Ortiz-Zambrano et al., 2025; Nohejl et al., 2026), while their feature-based nature enables controlled ablation analyses.

Second, we examine whether the observed effects persist in a stronger hybrid architecture that combines contextual predictions from a pretrained transformer language model with engineered features. This design is motivated by the design employed by Ortiz-Zambrano et al. (2025), who showed that manually engineered linguistic features complement contextual representations from pretrained language models in English and Spanish LCP. Their results indicate that hybrid models can outperform purely end-to-end approaches, particularly in multilingual settings. Accordingly, we use a `LightGBM` meta-regressor stacked on predictions from `XLM-RoBERTa-base` (Mosquera, 2021), which is the BEA shared-task baseline model. This architecture leverages the predictive capacity of a pretrained transformer model while retaining the ability to ablate the contribution of the feature groups introduced above.

We obtain the `RoBERTa` predictions using nested cross-validation to prevent data leakage. For each of five outer folds, we partition the corresponding training data into four inner folds. Using the HuggingFace `Trainer`,[4] we fine-tune `RoBERTa` on three inner folds and generate out-of-fold predictions for the remaining fold, repeating this procedure across all four splits. The resulting out-of-fold predictions provide a complete, leakage-free training signal for the `LightGBM` meta-regressor. Parallel instances of the same vocabulary item across the three source languages are always assigned to the same fold in both the outer and inner splits. Fine-tuning follows the hyperparameters and contextual input format of the BEA shared-task baseline: the L1 source word, the L1 source context sentence, the first letter of the English target word followed by underscores for its remaining characters as presented as a length cue to the learner, and the full English target word. See Appendix E.1 for an overview of both the `RoBERTa` and `LightGBM` hyperparameters we employ in our model training.

In parallel, we fine-tune a final round of `RoBERTa` models on the complete outer training partition and use them to predict the held-out outer test folds. The `LightGBM` models are trained on the out-of-fold predictions and engineered features extracted exclusively from the outer training data (see Ap-

[4] `https://huggingface.co/docs/transformers/en/main_classes/trainer`, last accessed 2026/10/30.

Config	Full	German	Spanish	Mandarin
B	.786 | 1.179	.786 | 1.193	.772 | 1.265	.804 | 1.071
E	.352 | 1.661	.429 | 1.608	.346 | 1.755	.266 | 1.617
^{L1}	.003 | 1.775	-.017 | 1.778	-.014 | 1.869	-.007 | 1.671
^{L1}C	.512 | 1.532	.466 | 1.583	.459 | 1.670	.616 | 1.321
^{L1}S	.498 | 1.544	.532 | 1.511	.513 | 1.607	.436 | 1.510
^{L1}E	.367 | 1.651**	.438 | 1.599**	.356 | 1.747**	.280 | 1.604**
^{L1}CS	.640 | 1.366	.651 | 1.352	.633 | 1.447	.635 | 1.293
^{L1}CE	.601 | 1.422**	.608 | 1.416**	.565 | 1.547**	.635 | 1.294**
^{L1}SE	.528 | 1.510**	.560 | 1.476**	.547 | 1.567**	.462 | 1.486**
^{L1}CSE	.659 | **1.337****	.669 | **1.323****	.657 | **1.411****	.649 | **1.273****

Table 2: Performance scores (Pearson r on the left-hand side; RMSE on the right-hand side) excluding baseline RoBERTa predictions for all models except B, which is the RoBERTa baseline predictions themselves. Performance gains of underlined configurations were tested for significance with paired one-sided sign-flip tests (10,000 permutations; $* = p < .05$, $** = p < .01$) against their corresponding configurations excluding E. Holm–Bonferroni correction was applied across all tests.

pendix E.2). At test time, they receive the final RoBERTa predictions for the outer test folds together with the corresponding engineered features.

We repeat this procedure independently for all five outer folds. Figure 5 in Appendix E.2 illustrates the inner–outer cross-validation procedure and the final meta-regressor training. Finally, we aggregate predictions across the outer test folds to compute overall and L1-specific performance metrics, as well as to conduct a sliding-window analysis of the NLD between source–target word pairs.

5.2 Evaluation

Following Felice and Skidmore (2026), our main evaluation metric is the root mean squared error (RMSE) between the predicted and gold complexity scores, which was used as the official evaluation metric in the BEA Shared Task. We additionally report Pearson r to quantify the strength of the linear relationship between predicted and observed complexity values. We report numerical performance scores across the entire dataset and the different L1 subsets individually. Following the BEA shared-task evaluation scheme, our visual analyses focus solely on the individual L1 subsets.

5.3 Experimental Setup

To answer the research questions introduced in Section 1, we conduct four experiments. EXPERIMENT 1 addresses RQ1 by evaluating the predictive potential of common lexical (C), surface-similarity (S), and etymology-based (E) feature groups, both individually and in all possible combinations. We additionally assess how their performance compares with that of the BEA shared-task RoBERTa baseline (B). To account for L1-specific complexity-score distributions (Skidmore et al., 2025), we include an explicit L1 feature in our analysis, denoted by the superscript ^{L1}.

EXPERIMENT 2 addresses RQ2 by examining the overlap of correct predictions produced by the individual components B^{L1}, ^{L1}C, ^{L1}S, and ^{L1}E. EXPERIMENT 3 addresses RQ3 through both a performance analysis across NLD windows and a candidate-adjusted enrichment analysis of exclusively correct predictions across NLD windows. Together, these analyses test whether the feature groups capture distinct signals and identify NLD ranges where these signals are most informative.

Finally, EXPERIMENT 4 addresses RQ4 by repeating the feature-group and NLD-window analyses from EXPERIMENT 1, with the RoBERTa baseline predictions (B) included in every configuration.

6 EXPERIMENT 1: Feature-Group Performance and Combinations Excluding Transformer Baseline

Setup. EXPERIMENT 1 evaluates the predictive potential of the three engineered feature groups in isolation. We train LightGBM models using E alone, L1 alone, and L1 combined with every possible subset of C, S, and E: ^{L1}C, ^{L1}S, ^{L1}E, ^{L1}CS, ^{L1}CE, ^{L1}SE, and ^{L1}CSE. We report the performance of the transformer baseline, B, as a reference point.

For each configuration containing E, we test whether its RMSE is significantly lower than that of the otherwise identical configuration without E. Thus, we compare ^{L1}E with ^{L1}, ^{L1}CE with ^{L1}C, ^{L1}SE with ^{L1}S, and ^{L1}CSE with ^{L1}CS.

Inclusion of an L1 Feature. Adding L1 as an additional component to the feature groups is useful as it allows the model to account for systematic differences in score distributions between the three L1 subsets, as illustrated by the slight performance improvements of ^{L1}E over E (see Table 2). We therefore include L1 as a regular feature component.

Results. As Table 2 illustrates, etymology features improve every matched feature configuration. Across the full dataset and the three L1 subsets, all 16 paired comparisons between models with and without E yield significant RMSE reductions after Holm correction. ^{L1}CSE is consistently

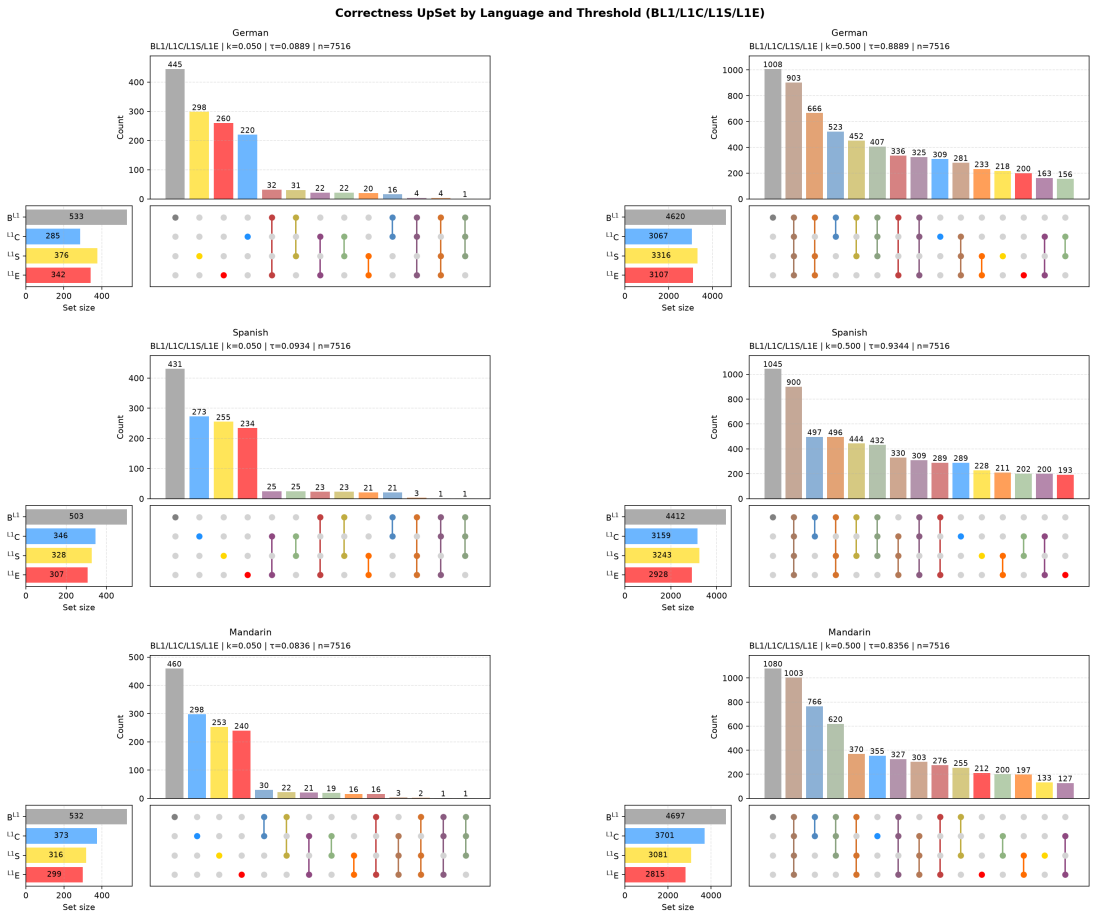


Figure 1: UpSet plots depicting the aggregate number of correct predictions per feature-group component (set size). The plots further quantify the overlap in model performance, contrasting the proportion of instances correctly classified by a singular model against those recognized by multiple models across the two evaluation thresholds.

the strongest configuration, achieving the lowest RMSE among the non-transformer models on both the pooled data and each L1 subset. Thus, although etymological information alone is insufficient to match the performance of common lexical features or the transformer baseline, it consistently complements both common and surface-similarity features.

While, by itself, the etymology group is the weakest feature group in terms of RMSE, it is clearly predictive: $^{L1}\mathsf{E}$'s Pearson r ranges from .280 for Mandarin to .356 for Spanish and .438 for German. This order is consistent with the greater lexical similarity that English shares with German and Spanish, which provides more opportunities for etymological relationships to be informative.

7 Experiment 2: Overlap of Correct Predictions

Setup. The aggregate results in Experiment 1 establish that the feature groups differ in predictive performance, but they do not reveal whether the models succeed on the same instances. To address RQ2, we therefore compare the overlap of correct predictions made by the individual base configurations: An L1-enriched transformer baseline B^{L1}, and the feature-only models $^{L1}\mathsf{C}$, $^{L1}\mathsf{S}$, and $^{L1}\mathsf{E}$.

Because the target variable is continuous, correctness must be operationalized using an error tolerance. For each L1 subset l, we define a prediction as correct if its absolute error does not exceed a threshold τ based on the standard deviation σ_l of the gold complexity scores in that subset. We report a strict threshold and a lenient threshold, $\tau \in \{0.05\sigma_l, 0.5\sigma_l\}$, to identify highly accurate predictions and assess whether the observed pattern remains when a broader range of predictions is considered acceptable.

We visualize the resulting intersections using UpSet plots (Lex et al., 2014; Figure 1). These plots show both the total number of correct predictions per configuration and the number of items

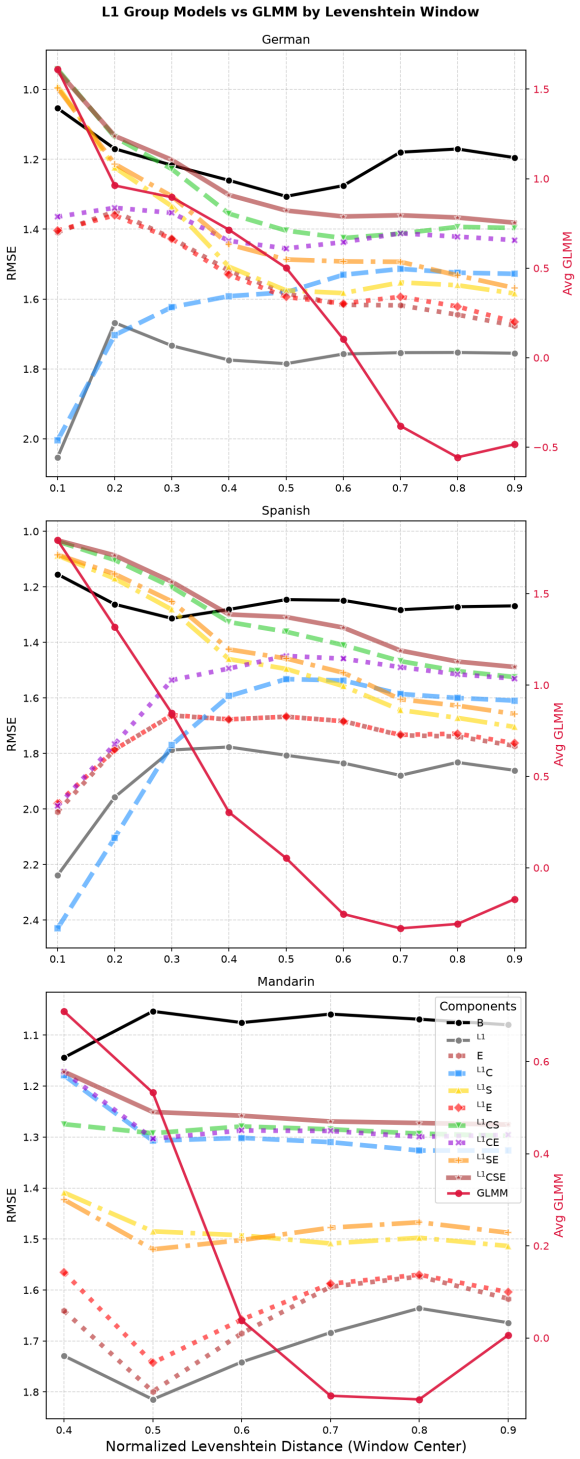


Figure 2: Sliding-window performance analysis of feature-group configurations across NLD windows, visualizing each feature group's predictive potential in isolation and in all possible combinations, excluding the transformer baseline predictions.

predicted correctly by each possible combination of models. Of particular interest are the single-model intersections, which represent instances correctly predicted by only one configuration.

Results. Figure 1 shows that none of the feature-group models is redundant with the transformer baseline or with the other feature-group models. Across all three L1 subsets and at both thresholds, each configuration produces instances that no other configuration predicts correctly.

As expected, the stricter threshold yields relatively little overlap between configurations: Only a small number of instances are predicted with very high accuracy by multiple models. Under the lenient threshold, overlap increases substantially because more predictions qualify as correct. Crucially, however, all models retain exclusive correct predictions under both definitions of correctness.

8 Experiment 3: NLD-Specific Performance and Enrichment

Setup 3A. To probe whether the usefulness of the feature groups differs depending on the surface similarity of the source–target word pair, we first conduct a sliding-window analysis of RMSE performance across NLD, scaled by maximum length. We calculate RMSE for overlapping windows of width 0.2, ranging from 0.0–0.2 to 0.8–1.0 in increments of 0.1. Alongside the RMSE curves, we plot the Generalized Linear Mixed Model-derived complexity score on a second y-axis to illustrate how vocabulary difficulty varies with the surface similarity between word pairs.

Results 3A. As Figure 2 shows, for German and Spanish, surface-similarity features are most useful at low NLD values, where source and target words are visibly similar. Their relative contribution declines as NLD increases. In contrast, models containing etymology features show their clearest advantage in intermediate NLD ranges, approximately 0.3–0.8. In these windows, they regularly outperform otherwise comparable models without E, particularly those relying on similarity features alone. At high NLD values, common lexical features become increasingly competitive and often provide the strongest signal, plausibly because lexical properties such as word frequency are much more informative and reliable predictors when cross-linguistic form similarity is low.

As expected, the pattern is less pronounced for Mandarin. Few Mandarin–English pairs occur below an NLD of 0.4; windows below this range therefore contain too few instances for reliable interpretation (see Appendix F). Across the better-supported windows, etymology features remain

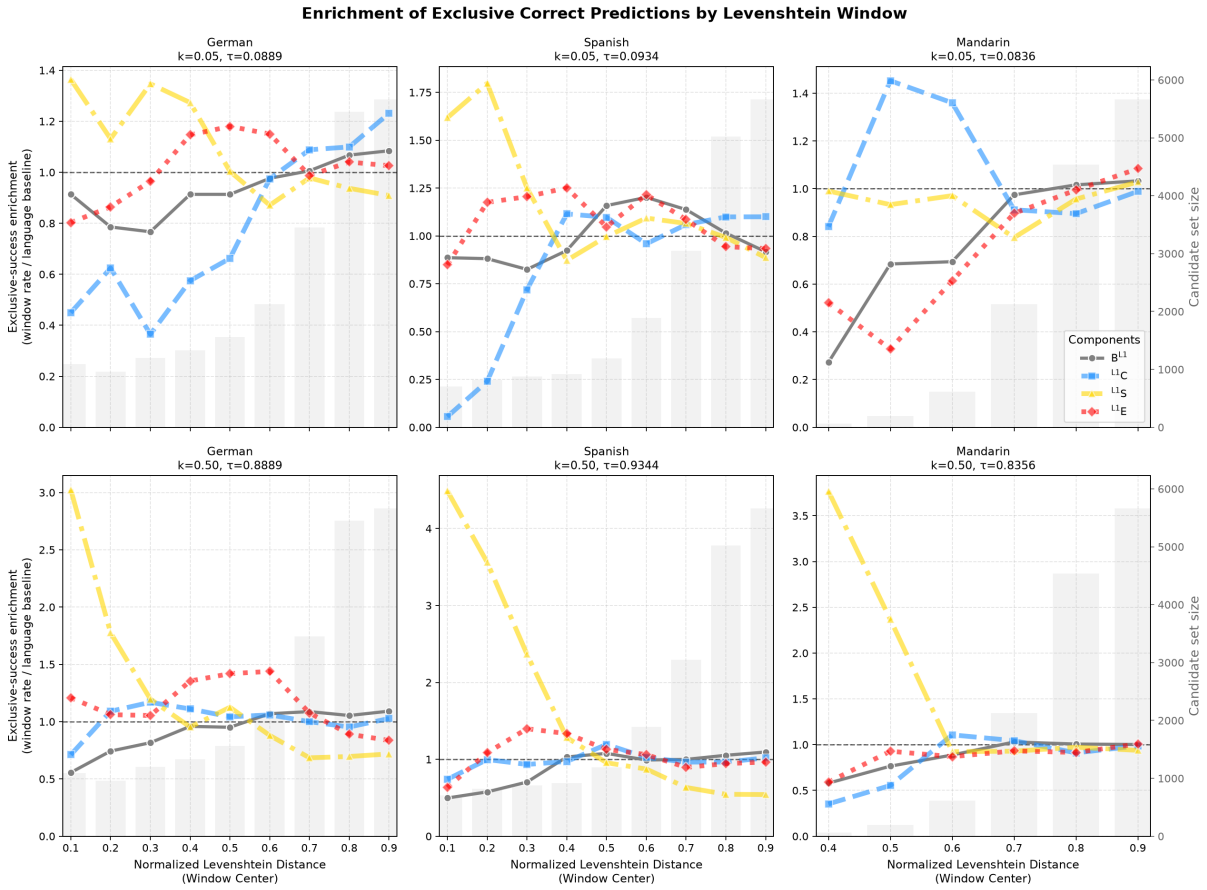


Figure 3: Candidate-adjusted enrichment analysis of exclusive correct predictions over NLD windows, visualizing each feature group's specific operational niche across L1s under $\tau \in \{0.05\sigma_l, 0.5\sigma_l\}$.

beneficial overall, especially in combination with common lexical features. However, common lexical features are generally more competitive than either similarity or etymology features, and the NLD-specific trends are much less distinct than for German and Spanish. This result is consistent with the observed lower number of surface-similar word pairs and the expected lower number of etymologically related word pairs in the Mandarin–English dataset.

Setup 3B. As mentioned above, the underlying dataset exhibits a natural skew, containing a disproportionately high volume of word pairs with higher NLDs (see Appendix F). Because of this imbalance, analyzing the raw distribution of correct predictions obscures a feature's true operational strengths by conflating model performance with candidate density per window. To address this issue and accurately isolate the specific operational niche of each feature set, we additionally calculate the candidate-adjusted enrichment of exclusive correct predictions per base feature model configuration (B^{L1}, $^{\mathsf{L1}}\mathsf{C}$, $^{\mathsf{L1}}\mathsf{S}$, $^{\mathsf{L1}}\mathsf{E}$). This metric adjusts for both local candidate density and the model's language-specific overall exclusive-success rate.

For each model configuration m, NLD window w, and source language l, we define the candidate-adjusted enrichment of exclusive correct predictions as:

$$E_{m,w,l} = \frac{U_{m,w,l}/N_{w,l}}{U_{m,l}/N_l}, \tag{1}$$

where $U_{m,w,l}$ is the number of predictions that are correct only for model m in window w and language l, $N_{w,l}$ is the number of candidate pairs in that window, $U_{m,l}$ is the total number of exclusive correct predictions made by m for language l, and N_l is the total number of instances for that language. As in EXPERIMENT 2, we calculate correctness under both the strict and lenient error thresholds. An enrichment value above 1 indicates that a model's exclusive correct predictions occur more often in a given NLD window than would be expected from its overall exclusive-success rate.

Config	Full	German	Spanish	Mandarin
B	.786 | 1.179	.786 | 1.193	.772 | 1.265	.804 | 1.071
BE	.779 | 1.114**	.779 | 1.116**	.763 | 1.209**	.798 | 1.007**
B^{L1}	.779 | 1.113	.780 | 1.113	.764 | 1.207	.797 | 1.010
$B^{L1}C$	.781 | 1.110	.779 | 1.116	.766 | 1.203	.801 | 1.002
$B^{L1}S$	.782 | 1.108	.780 | 1.114	.773 | 1.187	.794 | 1.015
$B^{L1}E$	.779 | 1.113	.780 | 1.114	.763 | 1.209	.798 | 1.008
$B^{L1}CS$	.792 | 1.084	.791 | 1.089	.786 | 1.157	.802 | 1.000
$B^{L1}CE$	.785 | 1.101**	.783 | 1.107*	.770 | 1.194*	.804 | **0.994***
$B^{L1}SE$	.784 | 1.103**	.783 | 1.108	.775 | 1.183	.796 | 1.011
$B^{L1}CSE$	.796 | **1.075****	.796 | **1.078****	.790 | **1.148****	.804 | **0.994***

Table 3: Performance scores (Pearson r on the left-hand side; RMSE on the right-hand side) including RoBERTa baseline predictions. B is the RoBERTa baseline predictions themselves. Performance gains of underlined configurations were tested for significance with paired one-sided sign-flip tests (10,000 permutations; $* = p < .05$, $** = p < .01$) against their corresponding configurations excluding E. Holm–Bonferroni correction was applied within hypothesis families (separated in the table).

Results 3B. The enrichment analysis identifies distinct NLD-specific niches for the feature groups (Figure 3). For German and Spanish, ^{L1}S is most strongly enriched at low NLD values, approximately 0.0–0.4. This result is expected: When source and target forms are visibly similar, surface-based measures provide a direct cue to cross-linguistic relatedness. As NLD increases, however, the enrichment of ^{L1}S declines and frequently falls below its language-specific baseline rate.

In contrast, ^{L1}E exhibits its clearest enrichment in intermediate NLD ranges, approximately 0.3–0.7, for both German and Spanish. In these windows, etymology features account for a disproportionate share of instances that neither common or similarity features nor the transformer baseline predict correctly. This pattern supports our hypothesis that etymological information is particularly informative when etymological relationships remain present and potentially accessible to the learner, but are no longer sufficiently transparent to be captured reliably by surface similarity alone. For German, examples of correct predictions of word pairs that share an etymological relationship made exclusively by the ^{L1}E model include *Hüfte* 'hip', *schießen* 'shoot', and *Kirsche* 'cherry'; for Spanish, we identify word pairs such as *vidua* 'widow', *salchicha* 'sausage', and *igualdad* 'equality'. See Appendix G for a more exhaustive list.

At higher NLD values, approximately 0.7–1.0, common lexical features become more strongly enriched. In these cases, target-word properties such as frequency, length, syllable count, and POS provide a stable signal when neither surface similarity nor etymological relationship is informative. The transformer baseline also tends to be more reliable at the high end of the NLD spectrum and comparatively less enriched for low-NLD pairs, which suggests that it likely relies on cues distinct from the information captured by similarity and etymology features.

The Mandarin results differ from the German and Spanish patterns, as expected. At the lenient threshold, similarity features show some enrichment in the intermediate range, whereas at the strict threshold, they are generally not enriched. Common lexical features are particularly informative in the middle-to-higher NLD range, approximately 0.4–0.7. Etymology features show limited positive enrichment, primarily at the highest NLD values. Examples of exclusively correct predictions by ^{L1}E of items with etymological relationships include 伏特加 *fútèjiā* 'vodka', and 汉堡 *hànbǎo* 'hamburger'. These results should be interpreted cautiously because Mandarin–English pairs are concentrated at high NLD values and low-NLD windows contain few instances. Overall, etymology features remain useful and slightly outperform similarity features, but no clear enrichment trend emerges across the statistically well-supported windows.

Taken together, EXPERIMENTS 2 and 3 show that etymology features capture a distinct predictive signal. For German and Spanish, this signal is concentrated in the intermediate NLD range, where surface similarity has become less reliable and common lexical features have not yet become dominant. For Mandarin, the signal is less clearly localized across NLD ranges.

9 EXPERIMENT 4: Feature-Group Performance Analysis Including Transformer Baseline

Setup. EXPERIMENT 4 addresses RQ4 by repeating the analyses from EXPERIMENT 1 (4A) and EXPERIMENT 3A (4B) in the hybrid architecture described in Section 5.1. Every LightGBM configuration now receives the out-of-fold RoBERTa predictions, B, as an additional input feature.

We evaluate BE, B^{L1}, and all combinations of B^{L1} and the three feature groups: $B^{L1}C$, $B^{L1}S$, $B^{L1}E$, $B^{L1}CS$, $B^{L1}CE$, $B^{L1}SE$, and $B^{L1}CSE$. We again test configurations containing E against their otherwise

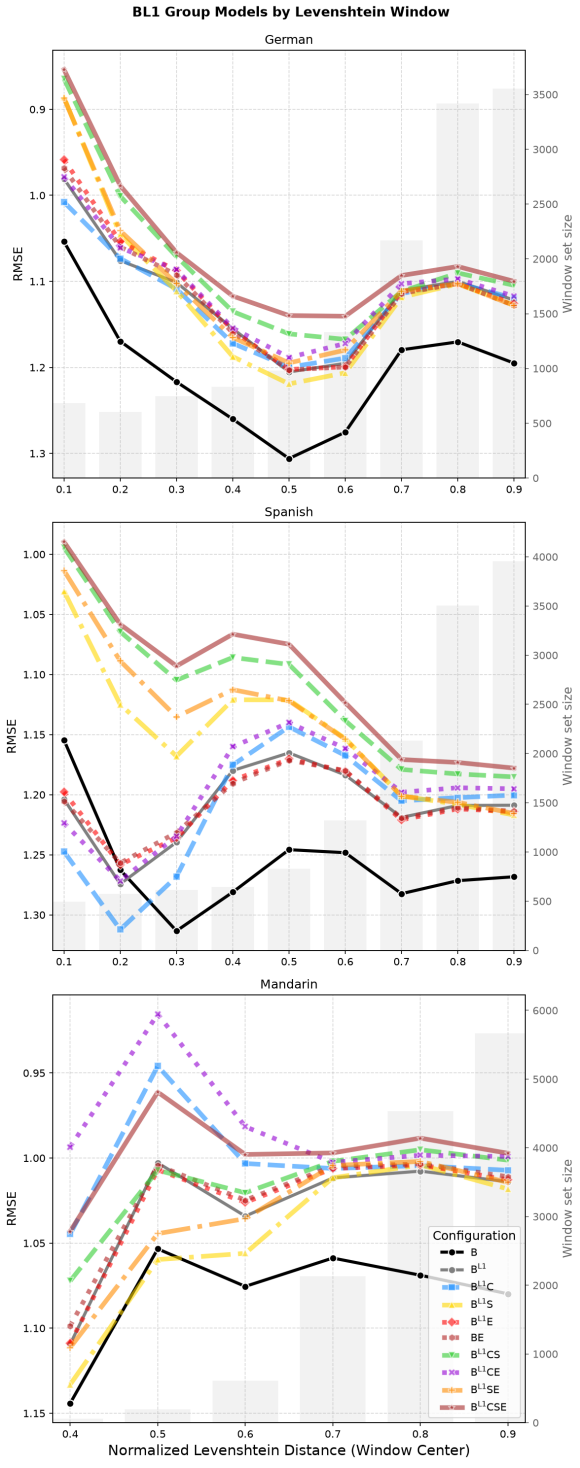


Figure 4: EXPERIMENT 4B: Sliding-window performance analysis of feature-group configurations across NLD windows, visualizing each feature group's impact on performance in all possible combinations, including the transformer baseline predictions.

identical counterparts without E. We also repeat the NLD sliding-window analysis.

Results. The hybrid results presented in Table 3 confirm that etymology features remain useful beyond the transformer baseline, although their effect is smaller than in the feature-only setting. Adding E to B yields a significant reduction in RMSE for the pooled data and for each L1 subset.

The best overall stacked configuration is $B^{L1}CSE$. It achieves the lowest RMSE on the pooled data as well as on the German and Spanish subsets. For Mandarin, $B^{L1}CSE$ and $B^{L1}CE$ obtain the same lowest RMSE. Crucially, $B^{L1}CSE$ significantly improves upon the corresponding configuration without etymology features, $B^{L1}CS$, for the full dataset and all three L1 subsets. Adding E to configurations containing common lexical features also produces significant improvements in every subset.

However, the effect of etymology features is not equally strong in all configurations. $B^{L1}E$ does not significantly improve upon B^{L1}, indicating that the etymology group alone offers limited additional value once the transformer prediction and L1 are already available. Likewise, $B^{L1}SE$ improves significantly only on the pooled data. These findings suggest that etymological information is most useful in the hybrid setting when combined with common lexical features ($B^{L1}CE$, $B^{L1}CSE$), which may provide a stable basis from which the model can exploit more sparse etymological signals.

The NLD sliding-window analysis broadly mirrors the feature-only results (Figure 4). For German and Spanish, similarity features provide the clearest gains at low NLD values. Etymology features contribute most visibly at intermediate distances: approximately 0.2–0.7 for Spanish and 0.3–0.8 for German. Within these ranges, $B^{L1}CSE$ often improves on $B^{L1}CS$, indicating that etymological information remains useful when surface similarity alone is insufficient.

For Mandarin, the patterns are less differentiated. In the relatively sparse lower-NLD windows, etymology features often outperform similarity features, but these results must be interpreted cautiously because they contain few observations. At high NLD values, where most Mandarin–English pairs occur, the three feature groups yield more comparable gains, while common lexical features remain particularly competitive. Overall, the stacked analysis confirms that etymology features can improve a strong transformer-based LCP system, especially in conjunction with common lexical features and for German and Spanish pairs at intermediate NLDs.

10 Discussion

Overall, our findings indicate that etymology features provide a useful and complementary signal for LCP in a multilingual context. Generally, etymology features exhibit moderate correlations with lexical complexity, suggesting that etymological relationships between source–target word pairs are associated with learners' ability to produce English vocabulary. This additional signal is reflected consistently in the experiments based exclusively on manually engineered features. Across all source languages, models that include etymology features significantly outperform their respective counterparts without these features, and the best performing model includes etymology features.

Furthermore, the effect of etymology differs across L1s. For German and Spanish, surface-similarity features are particularly informative for word pairs with low NLDs, where orthographic or phonological resemblance offers a clear indication of relatedness. As NLD increases into the intermediate range, however, etymology features regularly improve on similarity features. This pattern is consistent with our expectation that etymological information is especially useful when regular sound changes, language-specific adaptation, or orthographic divergence have weakened the surface similarity between etymologically related words. In such cases, surface similarity becomes a less reliable proxy for cognacy or shared borrowing histories, whereas etymology features can enrich the model with information about these relationships.

As expected, for Mandarin, the overall pattern is less pronounced. Although etymology features remain beneficial, similarity features are generally less informative, while common lexical features make a comparatively stronger contribution. This likely reflects the greater lexical and typological distance between Mandarin and English, which results in fewer source–target pairs with similar orthographic or phonological forms, as well as fewer candidates for shared etymological relationships. Correspondingly, we observe no similarly pronounced differences between NLD windows for Mandarin. Nevertheless, the positive contribution of etymology features suggests that historical borrowing relationships can still provide useful information even when surface similarity is limited.

These findings extend to the stacked architecture. Adding etymology features to `RoBERTa` predictions yields significant improvements in most configurations, confirming that etymological information is not fully captured by the transformer baseline predictions.

11 Conclusion

We investigated whether etymology-based features can improve LCP in a multilingual context beyond conventional surface-similarity measures. We derived nine features from Wiktionary etymology information for English target words and their German, Spanish, and Mandarin source-language equivalents, and evaluated their contribution both in models based on manually engineered features and in a stacked architecture combining these features with contextual `RoBERTa` predictions.

Adding etymology features significantly improved prediction performance in 29 of 36 matched comparisons ($\sim$81%), and the best-performing models always included etymology features. Our results show that these features distill an informative signal distinct from common lexical and surface-similarity features.

The contribution of etymology is strongest for German and Spanish source–target word pairs in the intermediate NLD range. For these pairs, surface-similarity-based features are less reliable predictors than at the lower end of the range, while common features do not fully capture the information encoded by etymological relationships. For Mandarin, etymology features remain useful overall, but the NLD pattern is less clearly localized. Etymology features may therefore be particularly valuable where historical relatedness has been obscured, but not eliminated, by sound change, phonological adaptation, or orthographic divergence.

Beyond improving LCP models, etymology features may benefit downstream EduNLP applications that rely on accurate estimates of vocabulary difficulty, such as lexical simplification, readability assessment, and adaptive language-learning systems. Future work should evaluate the approach on additional language pairs and investigate how it can be integrated into more advanced neural LCP architectures. It would also be valuable to examine how learners' individual linguistic backgrounds and explicit knowledge of etymological relationships affect the observed patterns. More broadly, the findings may have pedagogical implications for vocabulary selection and curriculum design, particularly where learning materials can draw attention to relevant cross-linguistic relationships.

Limitations

In this section, we raise awareness of a number of limitations of the present study.

First, our conclusions are constrained to the three source languages represented in the dataset: German, Spanish, and Mandarin. Other L1s may yield divergent patterns depending on their typological and lexical proximity to English, as well as their specific histories of language contact. The results for Mandarin warrant particular caution. While Wiktionary provides extensive etymological coverage for German and Spanish, only approximately 10% of the Mandarin source words in our data had corresponding etymological entries. For the remaining instances, `GPT-5.6-Luna` had to infer feature values from its own internalized knowledge. Since a manual inspection of the Mandarin items revealed that it is often non-native vocabulary that is covered by Wiktionary, which is the subject of our investigation, we do not anticipate this issue to fundamentally invalidate the utility of etymology features for Mandarin. However, it may compromise the reliability of individual feature values.

Our conclusions are also restricted to English as the target language and do not establish whether comparable gains would arise in the reverse direction, for example, for English-speaking learners of German, Spanish, or Mandarin. Nevertheless, we expect etymological information to be informative in other learning directions and, more generally, particularly so for language pairs with substantial genealogical relatedness or histories of language contact. For example, etymology features may be particularly promising for multilingual LCP involving languages such as Chinese, Japanese, Korean, and Vietnamese. This is because the latter three borrowed extensively from Chinese but subsequent divergent linguistic changes often obscure these relationships when comparing surface forms alone. These expectations require empirical validation beyond the present shared-task configuration.

Second, the reliability of our feature data is inherently tied to our extraction methodologies. In this study, we relied exclusively on `GPT-5.6-Luna` to parse Wiktionary entries. A manual inspection of a sample of the generated feature values revealed occasional divergences between human judgment and the model's outputs. Specifically, we observed inconsistencies in the `lca_language` feature, where the model struggled to accurately differentiate between Proto-West-Germanic and Proto-Germanic. Furthermore, we identified conflations of `same_proto_word` and `same_proto_root`, as well as failures to detect joint borrowings (`borrowed_from_third`). Future research must prioritize a more rigorous evaluation of LLM-generated features and compare multiple LLMs to ascertain whether the signal distilled by etymology features can be further denoised.

Third, methodological constraints in our preprocessing pipeline may partially account for the comparatively weak contribution of similarity features for Mandarin. To compute orthographic similarity, Chinese characters were romanized, tone marks and whitespace were removed, and the resulting pinyin forms were compared with English orthography. This representation may inadequately capture the phonological, orthographic, or lexical relationships that are salient to Mandarin-speaking learners. Subsequent investigations should explore similarity measures derived directly from Chinese characters, phonological representations, or linguistically more sophisticated transliteration protocols.

Fourth, at a theoretical level, our etymology features should be conceptualized as proxies for predictive signals within the data, rather than as direct reflections of learners' mental lexicons or their subjective experiences of lexical complexity. Although these features enhance predictive performance, our experimental design cannot establish whether learners explicitly recognize the encoded etymological relationships, which aspects of these relationships drive the observed effects, or whether a more direct operationalization of the same signal is feasible. Future work should therefore aim to isolate the relevant signal more accurately and develop targeted metrics or alternative feature representations. Moreover, as the authors of the Knowledge-Based Vocabulary Lists (KVL) project, which is the source of the BEA shared-task dataset, point out themselves (Schmitt et al., 2022),

> the level of knowledge which the [data] describes is the ability to accurately spell a lemma if the meaning is known. Research shows that this level of knowledge is relatively advanced, and is higher than receptive knowledge, [...] [b]ut the KVL cannot claim to represent full productive knowledge.

Thus, our claim that etymological information aids prediction can also only apply to this specific setup of measuring productive vocabulary knowledge.

Furthermore, L1 is only one learner-related factor relevant to lexical complexity and the personalization of learning solutions. Our study does not establish whether the relationship between etymological relatedness and lexical complexity changes systematically as learners' proficiency develops. Future work should therefore incorporate proficiency, for example represented by the CEFR level, alongside L1 and examine their interactions with etymology features. For pedagogical applications such as vocabulary selection, curriculum design, lexical simplification, and adaptive language learning systems, it is necessary to determine not only whether an item is easier for learners with a particular L1, but also at which proficiency levels such differences are pedagogically meaningful.

Fifth, the scope of our modeling choices remains limited. We evaluated only a single transformer baseline for contextual predictions alongside one stacking architecture. Future experiments should systematically compare distinct LLMs, non-LLM-based extraction techniques, contextual encoders, and hybrid-model architectures. Such comparative analyses are necessary to determine whether the observed performance gains are contingent upon specific implementation choices and whether the underlying etymological signal can be extracted with greater reliability.

Finally, while the etymology features demonstrate informational value, their performance gains are only marginal when integrated with the contextual transformer baseline. Given the comparatively high computational cost of LLM-based parsing, applying this procedure indiscriminately to all word pairs may prove inefficient. Etymology features may instead offer the greatest utility for word pairs in intermediate NLD ranges within lexically proximate languages (e.g., German and Spanish), where surface similarity provides less predictive power than at lower NLDs. Developing methodologies that either selectively apply etymological analyses in such contexts or automatically recover comparable signals without relying on LLMs could mitigate computational expenses and environmental impact while preserving the predictive benefits identified in this study.

Ethics Statement

When personalizing educational technologies based on learner characteristics, such as L1 backgrounds, it is important to ensure broad and equitable representation. However, this study is inherently constrained by the availability of existing datasets, which in the case of the BEA shared-task dataset exclusively features learners from high-resource languages, specifically German, Spanish, and Mandarin. Future research should prioritize the collection of data from more diverse linguistic communities to investigate whether these findings generalize across a more diverse population of learners.

Acknowledgments

We would like to thank Yuning Ding for her support in gaging the Mandarin data and our two anonymous reviewers for their valuable comments on potential future research directions.

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: We used several assistive generative AI tools while composing this paper. Most notably, we employed GitHub Copilot (with various LLM backbones) as an agentic coding assistant, and we used ChatGPT, Claude, and Gemini (in multiple versions) for brainstorming, literature searches, writing and editing assistance, and code and figure generation. We also consulted these models when interpreting our experimental results.

References

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–19, Rennes, France. LiU Electronic Press.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Exploiting the English vocabulary profile for L2 word-level vocabulary assessment with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 632–646, Vienna, Austria. Association for Computational Linguistics.

Karin Beijering, Charlotte Gooskens, and Wilbert Heeringa. 2008. Predicting intelligibility and per-

ceived linguistic distance by means of the levenshtein algorithm. *Linguistics in the Netherlands*, 25.

Yves Bestgen. 2026. SATLab at BEA 2026 shared task 1: Predicting the difficulty of English words for three L1 learners using primarily psycholinguistic features. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 985–991, San Diego, California, USA. Association for Computational Linguistics.

Marc Brysbaert, Guillermo Martínez, and Pedro Reviriego. 2025. Moving beyond word frequency based on tally counting: AI-generated familiarity estimates of words and phrases are an interesting additional index of language knowledge. *Behavior Research Methods*, 57:28.

Susanne E. Carroll. 1992. On cognates. *Interlanguage Studies Bulletin (Utrecht)*, 8(2):93–119.

David Crystal. 2008. *A Dictionary of Linguistics and Phonetics (6th ed.)*. Blackwell, Oxford.

Helen Engemann and Stefanie Radetzky. 2024. Cognate facilitation in child third language learners in a multilingual setting. *Languages*, 9(10).

Mariano Felice and Lucy Skidmore. 2026. Findings of the BEA 2026 shared task on vocabulary difficulty prediction for English learners. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 964–984, San Diego, California, USA. Association for Computational Linguistics.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online. Association for Computational Linguistics.

Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. 2023. Japanese lexical complexity for non-native readers: A new dataset. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 477–487, Toronto, Canada. Association for Computational Linguistics.

Elly Koutamanis, Gerrit Jan Kootstra, Ton Dijkstra, and Sharon Unsworth. 2024. Shared representations in cognate comprehension and production: An online picture naming and lexical decision study with bilingual children. *Applied Psycholinguistics*, 45(3):489–515.

Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992.

D. Löfflad and E. Kanli. 2026. Learner-centered automatic readability assessment: Leveraging hybrid models to consider l1. In *Proceedings of the 22. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*. In press.

Jonas Mayer Martins, Zhuojing Huang, Aaricia Herygers, and Lisa Beinborn. 2026. BoostedCats at BEA 2026 shared task 1: What makes a word hard to learn? modeling L1 influence on English vocabulary difficulty. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1047–1064, San Diego, California, USA. Association for Computational Linguistics.

Gerard de Melo. 2014. Etymological Wordnet: Tracing the history of words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1148–1154, Reykjavik, Iceland. European Language Resources Association (ELRA).

Noah-Manuel Michael, Anna Huelsing, and Andrea Horbach. 2026. GerVLPro: A CEFR-graded vocabulary list of L2 learners' productive vocabulary in German. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 11942–11959, Palma de Mallorca, Spain. ELRA Language Resource Association.

Alejandro Mosquera. 2021. Alejandro mosquera at SemEval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559, Online. Association for Computational Linguistics.

Adam Nohejl, Xuanxin Wu, Yusuke Ide, Maria Angelica Riera Machin, Yi-Ning Chang, and Hitomi Yanaka. 2026. Sakura at BEA 2026 shared task 1: What makes vocabulary difficult? In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1161–1178, San Diego, California, USA. Association for Computational Linguistics.

Kai North and Marcos Zampieri. 2023. Features of lexical complexity: Insights from L1 and L2 speakers. *Frontiers in Artificial Intelligence*, Volume 6.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

Jenny A. Ortiz-Zambrano, César H. Espín-Riofrío, and Arturo Montejo-Ráez. 2025. Deep encodings vs. linguistic features in lexical complexity prediction. *Neural Computing and Applications*, 37(3):1171–1187.

Jenny A. Ortiz-Zambrano and Arturo Montejo-Ráez. 2020. Overview of ALexS 2020: First workshop on lexical analysis at SEPLN. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, volume 2664 of *CEUR Workshop Proceedings*, Málaga, Spain. CEUR-WS.org.

Agnieszka Otwinowska and Jakub M. Szewczyk. 2019. The more similar the better? factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism*, 22(8):974–991.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. Adaptive complex word identification through false friend detection. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 192–200, New York, NY, USA. Association for Computing Machinery.

Vassili Philippov, Dmitrii Andreev, Pavel Katunin, and Anton Nikolaev. 2026. Glite at BEA 2026 shared task 1: Holistic difficulty models dominate, feature engineering closes the gap in L1-aware vocabulary difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1091–1105, San Diego, California, USA. Association for Computational Linguistics.

Santiago Robaina, Aiala Rosá, and Luis Chiruzzo. 2026. RETUYT-INCO at BEA 2026 shared task 1: Feature-enriched mDeBERTa for word difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1113–1118, San Diego, California, USA. Association for Computational Linguistics.

Job Schepens, Ton Dijkstra, and Franc Grootjen. 2012. Distributions of cognates in europe as based on levenshtein distance. *Bilingualism: Language and Cognition*, 15(1):157–166.

Job Schepens, Roeland van Hout, and T. Florian Jaeger. 2020. Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, 194:104056.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2022. *Knowledge-based Vocabulary Lists: User Manual: A Supplementary Resource*. British Council.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2024. *Knowledge-based Vocabulary Lists*. University of Toronto Press, Toronto.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Lucy Skidmore, Mariano Felice, and Karen Dunn. 2025. Transformer architectures for vocabulary test item difficulty prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 160–174, Vienna, Austria. Association for Computational Linguistics.

Anaïs Tack. 2021. Mark my words! On the automated prediction of lexical difficulty for foreign language readers. PhD Thesis.

Evy Woumans, Robin Clauws, and Wouter Duyck. 2021. Hands down: Cognate effects persist during written word production. *Frontiers in Psychology*, 12:647362.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France. European Language Resources Association.

A Abbreviations

Table 4 provides an overview of the abbreviations used in this paper.

B Feature Value Sets

Tables 5, 6, and 7 illustrate the features we explored in this study and their value ranges.

B	baseline
BEA	Workshop on Innovative Use of NLP for Building Educational Applications
C	common lexical (features/models)
CEFR	Common European Framework of Reference for Language
E	etymology (features/models)
EduNLP	educational NLP
GPT	Generative Pre-trained Transformer
IPA	International Phonetic Alphabet
KVL	Knowledge-Based Vocabulary Lists
LCP	Lexical Complexity Prediction
LightGBM	Light Gradient-Boosting Machine
LLM	Large Language Model
L1, ^{L1}	first language; source language(-based feature/models)
L2	second language; target language
NLD	Normalized Levenshtein Distance
NLP	Natural Language Processing
POS	part of speech
RMSE	root mean-squared error
RoBERTa	Robust Bidirectional Encoder Representations from Transformers
RQ	research question
S	surface-similarity (features/models)
TF-IDF	Term Frequency-Inverse Document Frequency
XLM-RoBERTa	Cross-Lingual Language Model with RoBERTa Architecture

Table 4: Abbreviations; alphabetically ordered.

Feature	Values
`zipf_frequency`	$[1.17, 7.17]$
`target_length`	$\{1, 2, \dots\}$
`num_syllables`	$\{1, 2, \dots\}$
`pos_tag`	noun verb adjective adverb determiner preposition number not-no misc

Table 5: Sets of possible values for common lexical features.

Feature	Values
`char_edit_dist`	$\{0, 1, 2, \dots\}$
`char_edit_dist_norm`	$[0, 1]$
`char_edit_dist_geo_norm`	$[0, \infty)$
`phon_edit_dist`	$\{0, 1, 2, \dots\}$
`phon_edit_dist_norm`	$[0, 1]$
`phon_edit_dist_geo_norm`	$[0, \infty)$
`jaro_dist`	$[0, 1]$
`jaro_winkler_dist`	$[0, 1]$
`soundex_dist`	$\{0, 1, 2, 3, 4\}$
`metaphone_dist_norm`	$[0, 1]$
`char_ngram_tfidf_sim`	$[0, 1]$

Table 6: Sets of possible values for similarity features.

C Etymology Feature Extraction Prompt

You are an expert historical linguist. You will be presented with a source-target word pair (possible source languages: German, Spanish, Mandarin Chinese; target language: English). You must return a JSON object (dictionary) corresponding to the features described below. If available, you will receive the Wiktionary etymology entries for the source and target words for additional context. If there is no Wiktionary information provided, please answer to the best of your ability. Since the POS tag of the source word is not given, we also present you with an example of the source word in context. If there are multiple etymology entries for a source word, choose the one that matches the POS and meaning you infer from the source context as your reference.

Word Relatedness Features

Feature 1: Lowest Shared Genetic Proto-Language

Description: Identifies the youngest reconstructed ancestral language that both words naturally evolved from. This strictly measures genetic lineage, ignoring contact, borrowing, or donor languages.

Type: Categorical

Feature	Values
lca_language	Proto-West-Germanic Proto-Germanic Proto-Indo-European None
borrowed_from_source	$\{0, 1\}$
borrowed_from_target	$\{0, 1\}$
borrowed_from_third	$\{0, 1\}$
english_word_origin	Germanic Romance Sinitic Other
origin_matches_source_family	$\{0, 1\}$
same_proto_word	$\{0, 1\}$
same_proto_root	$\{0, 1\}$
transmission_path	Inherited Direct Indirect

Table 7: Sets of possible values for etymology features.

Values:

- Proto-West-Germanic: Shared genetic ancestor at the West Germanic level.
- Proto-Germanic: Shared genetic ancestor at the broader Germanic level.
- Proto-Indo-European: Shared genetic ancestor at the Proto-Indo-European level.
- None: Applicable if the words do not share a reconstructed genetic proto-language, OR if their only shared history is via borrowing (e.g., both borrowed from French or Latin).

Further Comments: None.

Feature 2: Borrowed from Source Language

Description: Indicates whether the English target word was directly borrowed from the specified source language.
Type: Binary
Values:

- 0: English target word was not borrowed from the source language.
- 1: English target word was directly borrowed from the source language (e.g., English *kindergarten* from German).

Further Comments: None.

Feature 3: Borrowed from Target Language

Description: Indicates whether the source word was directly borrowed from English.
Type: Binary
Values:

- 0: Source word was not borrowed from English.
- 1: Source word was directly borrowed from English (e.g., German *Computer*, Spanish *fútbol*).

Further Comments: None.

Feature 4: Borrowed from Same Third Language

Description: Indicates whether both the source word and the English target word were borrowed from the same non-source, non-English donor language.
Type: Binary
Values:

- 0: No evidence that both words were borrowed from the same third language.
- 1: Both words were borrowed from the same third language (e.g., German *flexibel* and English *flexible* both from Latin/French transmission).

Further Comments: This feature is independent from direct borrowing between source and target (Features 2 and 3).

Feature 5: Immediate Taxonomic Branch of English Word

Description: Classifies the broad language branch or family of the immediate predecessor from which modern English acquired the word. This does not ask for the specific language, but rather the taxonomic category of the immediate donor or ancestor.
Type: Categorical
Values:

- Germanic: Entered from Old English, Old Norse, Dutch, German, etc.
- Romance: Entered from French, Latin, Spanish, Italian, etc.
- Hellenic: Entered from Greek.

- `Sinitic`: Entered from Mandarin, Cantonese, etc.
- `Other`: Any other taxonomic branch not covered above.

Further Comments: None.

Feature 6: Origin Matches Source Language Family?

Description: Determines if the immediate linguistic origin of the English target word aligns with the broader language family of the source word.
Type: Binary
Values:

- `0`: English target word origin does not match source language family.
- `1`: English target word origin matches source language family.

Further Comments: Assign `1` if the English target word's immediate taxonomic branch (assigned in Feature 5) matches the broader language family of the source language. Specifically: if the source language is German, assign `1` if Feature 5 is 'Germanic'; if the source language is Spanish, assign `1` if Feature 5 is 'Romance'; if the source language is Mandarin Chinese, assign `1` if Feature 5 is 'Sinitic'. Otherwise, assign `0`.

Feature 7: Descendant of the Same Proto-Word?

Description: Evaluates if both words descend from the exact same reconstructed ancestral lexical item at the youngest shared proto-language stage.
Type: Binary
Values:

- `0`: Source and target words do not descend from the same reconstructed ancestral lexical item. They may share a broader root, but are not descendants of the same proto-word.
- `1`: Source and target words are descendants of the same reconstructed ancestral lexical item at the youngest relevant shared proto-language stage.

Further Comments: "Proto-word" means a reconstructed lexical form, not merely a root or semantic concept. Two words count as sharing the same proto-word even if they diverged into different language branches. Direct cognacy does not require the modern forms to look similar. If a shared reconstructed lexical item exists, assign `1`. Examples: Spanish *padre* and English *father* both descend from Proto-Indo-European '$*ph_2t\acute{e}r$' (Assign 1). Spanish *pie* and English *foot* both descend from the exact same Proto-Indo-European proto-word '$*p\acute{o}ds$' (Assign 1).

Feature 8: Descendant of the Same Proto-Root?

Description: Evaluates if the words share a broader ancestral root or word family, specifically when they do not descend from the exact same reconstructed proto-word.
Type: Binary
Values:

- `0`: Source and target words do not share any reconstructable ancestral root or word family.
- `1`: Source and target words share a broader ancestral root or word family but are NOT descendants of the same reconstructed proto-word.

Further Comments: Assign `1` if the words share a historical root or word family but do NOT descend from the exact same reconstructed proto-word. F7 and F8 are mutually exclusive based on the youngest shared ancestor: if they share a proto-word, assign F7=1 and F8=0. If they only share a broader root, assign F7=0 and F8=1. Example: Spanish *juventud* and English *youth* both stem from the shared Proto-Indo-European root for 'young', but they do not descend from the exact same proto-word because the Italic and Germanic branches added completely different suffixes to that base root to form their respective nouns (F7=0, F8=1).

Feature 9: English Transmission Mechanism

Description: Describes the mechanical path of how the word reached Modern English, ignoring which languages were involved and focusing entirely on the complexity of the transmission history.
Type: Categorical
Values:

- `Inherited`: Passed down continuously into Modern English from its deepest ancestral roots with no borrowing events (0 hops).

- `Direct`: Borrowed directly into English from the ultimate source language with no intermediary languages (1 hop).
- `Indirect`: Passed through at least one intermediary language before reaching English (2+ hops).

Further Comments: None.

One-Shot Example

Input:
Spanish source word: padre
English target word: father

Source context:
Mi padre siempre me da buenos consejos, y mi madre también.

Wiktionary Info:
Source word in Spanish Wiktionary:
Del latín pater y este del protoindoeuropeo '$*ph_{\circ 2}t\acute{e}r$'. Compárese el protogermánico '$*fa\eth er$', el griego micénico pa-te, el griego clásico '$\pi\alpha\tau\acute{\eta}\rho$' (patér), el persa *pedar* o el sánscrito *pitr*.

Source word in English Wiktionary:
Inherited from Old Spanish padre ("father"), from Latin patrem, pater, from Proto-Italic '$*patēr$', from Proto-Indo-European '$*ph_2t\acute{e}r$'.

Target word in Spanish Wiktionary:
Del inglés medio fader, del inglés antiguo fæder, del protogermánico occidental '$*fader$', del protogermánico '$*fadēr$', del protoindoeuropeo '$*ph_{\circ 2}t\acute{e}r$'.

Target word in English Wiktionary:
Inherited from Middle English fader, from Old English fæder, from Proto-West Germanic '$*fader$', from Proto-Germanic '$*fadēr$', from Proto-Indo-European '$*ph_2t\acute{e}r$', possibly from '$*peh_2-$' + '$* - t\acute{e}r$'. Doublet of ayr, faeder, athair, padre, pater and père.

Output:

```
{
  "F1: Lowest Shared Genetic
 Proto-Language": "Proto-Indo-European",
  "F2: Borrowed from Source Language": 0,
  "F3: Borrowed from Target Language": 0,
  "F4: Borrowed from Same Third
  Language": 0,
  "F5: Immediate Taxonomic Branch of
  English Word": "Germanic",
  "F6: Origin Matches Source Language
  Family": 0,
  "F7: Descendant of the Same
  Proto-Word": 1,
  "F8: Descendant of the Same
  Proto-Root": 0,
  "F9: English Transmission Mechanism":
  "Inherited"
}
```

Your Task

Now, process the following word pair and return ONLY the JSON object. Do not include any markdown formatting, explanations, or text outside the JSON object.

Input:
{source_lang} source word: {l1_source}
English target word: {en_target}

Source context:
{source_context}

Wiktionary Info:
Source word in {source_lang} Wiktionary:
{source_etymology_l1}

Source word in English Wiktionary:
{source_etymology_en}

Target word in {source_lang} Wiktionary:
{target_etymology_l1}

Target word in English Wiktionary:
{target_etymology_en}

L1	Target	Source
German	93.9%	88.7%
Spanish	95.0%	83.2%
Mandarin	93.9%	10.1%

Table 8: Percentage of target and source words covered for etymological information in either the English or the respective L1 Wiktionary edition.

D Available Wiktionary Entries

Table 8 illustrates how many source and target words per L1 had Wiktionary etymology information available in either the English or the respective L1 Wiktionary edition.

Parameter	Value
Prediction head	regression
Maximum sequence length	256 tokens
Number of epochs	5
Training batch size	32
Learning rate	3×10^{-5}
Weight decay	0.1
Warmup ratio	0.1
Optimizer	AdamW
Random seed	667

Table 9: `XLM-RoBERTa-base` fine-tuning settings, using the HuggingFace `Trainer` (`transformers` 5.15.1, PyTorch 2.5.1) on a single GPU without gradient accumulation. Unlisted `TrainingArguments` take their `transformers` 5.15.1 defaults.

Parameter	Value
Objective	regression
Number of estimators	700
Learning rate	0.03
Number of leaves	63
Column sample by tree	0.9
Random seed	667

Table 10: `LightGBM` settings, using `LGBMRegressor` from `lightgbm` 4.7.0. Unlisted parameters take their `lightgbm` 4.7.0 defaults.

E Model Architecture

In this section, we provide supplementary materials concerning our model architecture.

E.1 Model Parameters

For reproducibility, we report the hyperparameters of the models we train (Tables 9 and 10).

E.2 Cross-Validation Logic

Figure 5 illustrates the inner and outer fold logic as well as the final `LightGBM` training.

F Number of Items per Levenshtein Window

Table 11 provides an overview of the number of source–target word pairs contained in each sliding NLD window per L1.

G Examples of Exclusively Correct Predictions by Etymology Model

Here, we present examples of exclusively correct predictions by the $^{\text{L1}}$E model at $\tau \in \{0.05\sigma_l, 0.5\sigma_l\}$ that exhibit an etymological relationship.

G.1 5% STD Threshold

We provide a selection of etymologically related source-target word pairs per L1, identified correctly exclusively by the $^{\text{L1}}$E model at $\tau = 0.05\sigma_l$.

German: Hüfte ‘hip’, schießen ‘shoot’, durch ‘through’, verdoppeln ‘double’, Sonnenschein ‘sunshine’, Mutterschaft ‘motherhood’, gegeben ‘given’, typisch ‘typically’, verboten ‘forbidden’, Französisch ‘french’, Aktivierung ‘activation’, komponieren ‘compose’, Kupfer ‘copper’, gesalzen ‘salted’, Fett ‘fat’, aktiv ‘actively’, identifizieren ‘identify’, Dank ‘thanks’, lachend ‘laughing’, rasten ‘rest’, Klassifizierung ‘classification’, Sonnenlicht ‘sunlight’, Soldat ‘soldier’, Delfin ‘dolphin’, Schande ‘shame’, Eisberg ‘iceberg’, frustrierend ‘frustrating’, geologisch ‘geologically’, romantisch ‘romantically’, Diskriminierung ‘discrimination’, Asche ‘ash’, statistisch ‘statistically’, Mystiker ‘mystic’, Handvoll ‘handful’, ungewaschen ‘unwashed’, Grammatik ‘grammar’, identisch ‘identical’, Wäsche ‘wash’, stressig ‘stressful’, Christ ‘christian’, unkontrolliert ‘uncontrolled’, statisch ‘static’, ethnisch ‘ethnic’, Mikrobiologie ‘microbiology’, psychologisch ‘psychological’, Adjektiv ‘adjective’, Therapeut ‘therapist’, Ethik ‘ethic’, harmonisch ‘harmonic’, national ‘nationally’, Psychologie ‘psychology’, homosexuell ‘homosexual’, Philosophie ‘philosophy’, Heterosexualität ‘heterosexuality’, Miniatur- ‘miniature’, evolutionär ‘evolutionary’, multifunktional ‘multifunctional’, Designer ‘designer’, Countdown ‘countdown’, Cheerleader ‘cheerleader’, Freak ‘freak’, Layout ‘layout’, Asteroid ‘asteroid’.

Spanish: viuda ‘widow’, salchicha ‘sausage’, ciudadanía ‘citizenship’, abordo ‘aboard’, segundo ‘secondly’, captura ‘catch’, incierto ‘uncertain’, comprometerse ‘promise’, tratar ‘treat’, tipo ‘type’, brillantemente ‘brilliantly’, rito ‘ritual’, redondeada ‘rounded’, química ‘chemical’, receta ‘recipe’, cartón ‘card’, alfabeticamente ‘alphabetically’, auditoría ‘audit’, demográficamente ‘demographically’, batalla ‘battle’, insatisfecho ‘unsatisfied’, independientemente ‘independently’, pó-

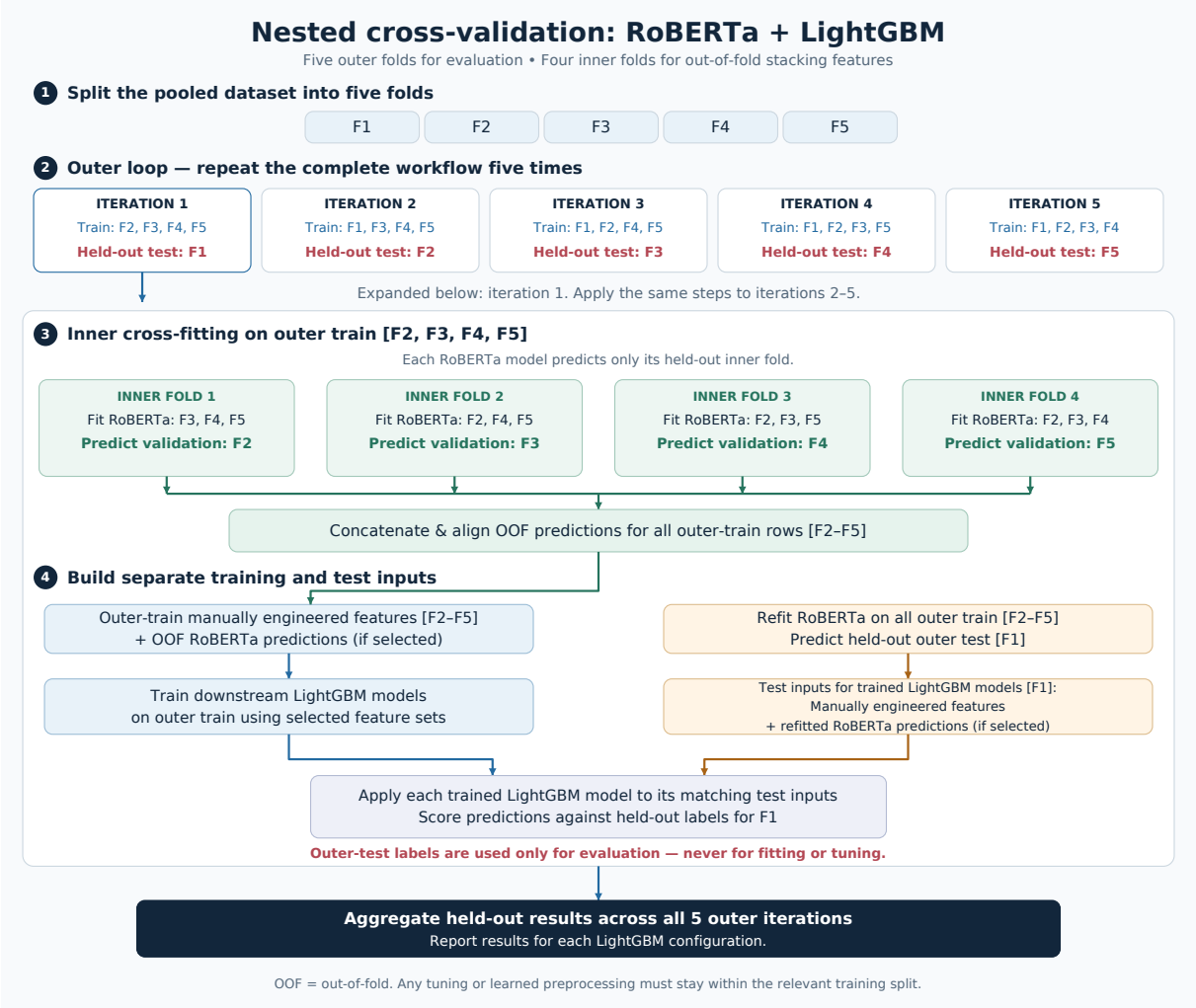


Figure 5: Illustration of our stacked RoBERTa–LightGBM model architecture and cross-validation loop.

quer ‘poker’, bárbaro ‘barbarian’, incontrolado ‘uncontrolled’, especialidad ‘speciality’, gimnasio ‘gymnasium’, contenido ‘content’, biografico ‘biographic’, instalador ‘installer’, instrumentalmente ‘instrumentally’, colonia ‘colony’, comparación ‘comparison’, insensible ‘insensitive’, estereotipo ‘stereotype’, inversor ‘investor’, intermedio ‘intermediate’, duplicar ‘duplicate’, cinematografía ‘cinematography’, rollo ‘roll’, adolecente ‘adolescent’, hipótesis ‘hypothesis’, microscopio ‘microscope’, hormona ‘hormone’, tradicionalista ‘traditionalist’, genética ‘genetics’, molécula ‘molecule’, celular ‘cellular’, vainilla ‘vanilla’, manipulador ‘manipulator’, propaganda ‘propaganda’, editorial ‘editorial’.

Mandarin: 伏特加 *fútèjiā* ‘vodka’, 卡拉OK *kǎlā OK* ‘karaoke’.

G.2 50% STD Threshold

We provide a selection of etymologically related source-target word pairs per L1, identified correctly exclusively by the ^{L1}E model at $\tau = 0.5\sigma_l$.

German: Kirsche ‘cherry’, schießen ‘shoot’, wöchentlich ‘weekly’, überhitzen ‘overheat’, Bohnen ‘bean’, unbekannt ‘unknown’, blutig ‘bloody’, reagieren ‘react’, halten ‘hold’, sollen ‘should’, aktivieren ‘activate’, Bruderschaft ‘brotherhood’, unzerbrechlich ‘unbreakable’, hausgemacht ‘homemade’, Weisheit ‘wisdom’, abgerundet ‘rounded’, Regentropfen ‘raindrop’, Älterer ‘elder’, verboten ‘forbidden’, kompliziert ‘complicated’, französisch ‘french’, Feuerwehrmann ‘fireman’, konzentriert ‘concentrated’, Streifen ‘strip’, gesalzen ‘salted’, Herzensbrecher ‘heartbreaker’, gestoppt ‘stopped’, hilflos ‘helpless’, beißen ‘bite’, lachend ‘laughing’, laut ‘loud’, Rezept ‘recipe’, Klassifizierung ‘classification’, unkompliziert ‘uncomplicated’, konzentrieren ‘concentrate’, hasserfüllt ‘hateful’, vibrieren ‘vibrate’, kontrollierbar ‘controllable’, unfreundlich ‘unfriendly’, Daten ‘data’, frustriert

	Window Center								
L1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
German	684	602	749	832	980	1332	2166	3418	3551
Spanish	492	574	613	642	828	1321	2131	3506	3955
Mandarin	0	2	17	60	191	612	2130	4535	5663

Table 11: Number of candidate items per Normalized Levenshtein Distance window in EXPERIMENTS 3A, 3B, 4B.

'frustrated', sauer 'sour', abgefallen 'fallen', Diskriminierung 'discrimination', dominieren 'dominate', ungewaschen 'unwashed', integrieren 'integrate', biografisch 'biographic', explodieren 'explode', Wäsche 'wash', Soziologie 'sociology', Kreativität 'creativity', Republikaner 'republican', exklusiv 'exclusive', Immunität 'immunity', Stock 'stick', Wagen 'wagon', Hallo 'hello', Subkultur 'subculture', Helfer 'helper', Flexibilität 'flexibility', mainstreamig 'mainstream', Elektron 'electron', Republik 'republic', Zigarette 'cigarette', Manuskript 'manuscript', Präsentation 'presentation', Leopard 'leopard', Exhibitionist 'exhibitionist', Astronaut 'astronaut', minimal 'minimal', Hunger 'hunger', Park 'park', Algebra 'algebra', Collage 'collage', Bronze 'bronze', bitter 'bitter'.

Spanish: jabón 'soap', sinfonía 'symphony', novedad 'innovation', igualmente 'equally', cubrir 'cover', solitario 'solo', igualdad 'equality', dirigido 'directed', esclavitud 'slavery', gráficamente 'graphically', sugerir 'suggest', elegido 'elected', tostadora 'toaster', sobrevivir 'survive', gravedad 'gravity', embajada 'embassy', boxeo 'boxing', cartón 'card', esquadrón 'squad', movilidad 'mobility', párrafo 'paragraph', ejercitar 'exercise', cocina 'cooking', champú 'shampoo', simplicficar 'simplify', multiplicar 'multiply', inseguridad 'insecurity', alfabético 'alphabetical', graduado 'graduated', decano 'dean', especialidad 'speciality', conectividad 'connectivity', aventurar 'venture', movimiento 'movement', publicidad 'publicity', gigante 'giant', calcio 'calcium', diccionario 'dictionary', sal 'salt', complejo 'complex', inspirado 'inspired', cristiano 'christian', ignoratne 'ignorant', radioactividad 'radioactivity', gimnástico 'gymnastic', óptica 'optics', incondicional 'unconditional', caramelo 'caramel', guitarista 'guitarist', legislación 'legislation', quántum 'quantum', tampón 'tampon'.

Mandarin: 伏特加 *fútèjiā* 'vodka', 汉堡 *hànbǎo* 'hamburger'. 伊拉克人 *yīlākè rén* 'Iraqi'.