BoostedCats at BEA 2026 Shared Task 1: What Makes a Word Hard to Learn? Modeling L1 Influence on English Vocabulary Difficulty

Jonas Mayer Martins **Zhuojing Huang** **Aaricia Herygers** **Lisa Beinborn**
University of Göttingen, Germany
firstname.lastname@uni-goettingen.de

Abstract

What makes a word difficult to learn, and how does the difficulty depend on the learner's native language? We computationally model vocabulary difficulty for English learners whose first language is Spanish, German, or Chinese with gradient-boosted models trained on features related to a word's familiarity (e.g., frequency), meaning, surface form, and cross-linguistic transfer. Using Shapley values, we determine the importance of each feature group. Word familiarity is the dominant feature group shared by all three languages. However, predictions for Spanish- and German-speaking learners rely additionally on orthographic transfer. This transfer mechanism is unavailable to Chinese learners, whose difficulty is shaped by a combination of familiarity and surface features alone. Our models provide interpretable, L1-tailored difficulty estimates that can be used to design vocabulary curricula.

Code repository and interactive demo

1 Introduction

Learning a second language begins with words. Developing an extensive vocabulary is essential for mastering grammar and achieving fluency (Schmitt and Schmitt, 2020; Nation, 2000). Yet not all words are equally hard to learn, and researchers of second-language (L2) acquisition, classroom practitioners, and creators of educational materials seek to better understand this phenomenon.

The difficulty of a word is partially due to factors common to all learners, that is, lexical properties that depend only on the target language itself (Peters, 2019). For example, common and concrete words are easier to acquire than rare and abstract ones (Ellis, 2002; Ellis and Beaton, 1993).

However, a learner's individual background—their first language (L1), in particular—plays a key role in word difficulty (Ringbom and Jarvis, 2009). For instance, a German speaker readily infers that the English word "sheep" is related to the cognate *Schaf*, whereas the Spanish *oveja* is not a cognate.[1] This orthographic bridge supports active as well as passive vocabulary competence. Conversely, false friends may lead to interfering transfer (Bensoussan and Laufer, 1984). For example, *bravo* in Spanish is close to its English cognate "brave", yet *brav* in German means "well-behaved". The same English word can thus be easy for one learner while difficult for another, depending on the relationship between the L1 and L2.

However, computational models of lexical difficulty often fail to take a learner's background into account. The SemEval2021 shared task on lexical-complexity prediction, for example, considers difficulty for unspecified annotators (Shardlow et al., 2021). Meanwhile, predictions of lexical difficulty tailored to individual learners are receiving increasing attention (Palmero Aprosio et al., 2020; North et al., 2023; Schmitt et al., 2021; Skidmore et al., 2025), yet studies that do include L1 effects often remain limited to a single language pair.

Approach and contributions. We address this gap by modeling English vocabulary difficulty separately for Spanish, German, and Chinese[2] L1 speakers, using the Knowledge-based Vocabulary Lists (KVL), a recent crowd-sourced corpus of lexical-difficulty tests (Schmitt et al., 2021; Skidmore et al., 2025), in the context of the closed track of the Building Educational Applications (BEA) 2026 shared task (Felice and Skidmore, 2026). We create four feature groups: the familiarity of a word based on exposure effects, its meaning based on semantic complexity and concreteness, its surface form based on orthographic complexity, and its transferability from a specific L1 based on string similarity. We train interpretable, gradient-boosted

[1] Instead, *oveja* is distantly related to "ewe", meaning female sheep, in English.

[2] In this article, we refer to Mandarin Chinese as Chinese.

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1047–1064

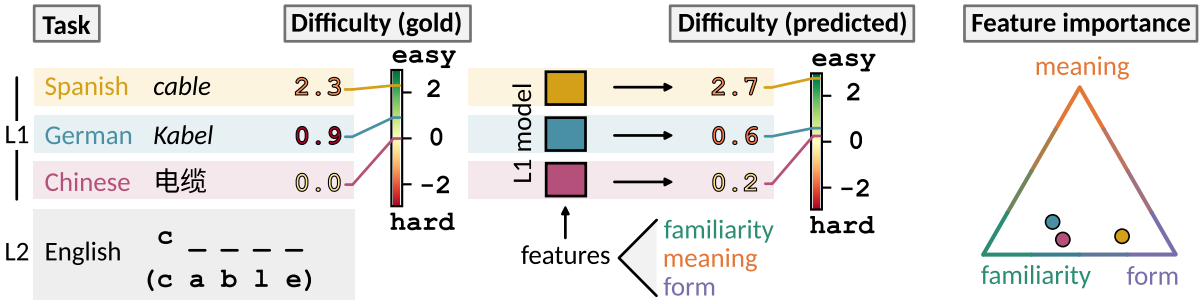

Figure 1: Illustration of the task and modeling setup. For each L1, learners translate a word (e.g., Kabel) from their L1 into English given the first letter (c) and a graphical hint indicating the number of characters. Item difficulty is estimated from aggregated responses (gold label). We train L1-specific models on feature groups (familiarity, meaning, and form = surface ∪ transfer). By evaluating the feature importance, we assign a relative importance in the prediction of a word to each feature group.

models on these features to predict L1-dependent lexical difficulty and trace their predictive power via Shapley values. Figure 1 illustrates the task format and our approach.

To showcase our results and support further research and practical use, we have created an interactive demo that extends our predictions to more than 4,000 to 7,000 additional English words per language, see Appendix A and Appendix B. This demo highlights the advantage of our computational modeling approach, which allows us to make predictions beyond the limited KVL data for a larger range of vocabulary and potentially more L1 backgrounds, providing tailored difficulty estimates applicable to curriculum design.

2 Related work

In this section, we discuss factors of word difficulty in a foreign language based on second-language-learning studies and computational models.

2.1 Lexical difficulty

Vocabulary knowledge poses a fundamental challenge to all second-language (L2) learners (Schmitt and Schmitt, 2020; Nation, 2000). However, the difficulty associated with a particular word varies by learner. Some words, for instance, represent intrinsically difficult concepts in the target language, while others are challenging or easy specifically due to a learner's L1 background.

L2-intrinsic difficulty. The more frequent a stimulus occurs, the more readily it tends to be recalled, and as such, frequent words are learned more easily because learners encounter them more often (Ellis, 2002; Peters, 2019). Educational curricula typically introduce frequent words first to enable learners to speak and understand the target language earlier by efficiently building vocabulary coverage (Nation, 2000).

Beyond frequency, other lexical properties are also important. For instance, orthographic complexity has been shown to predict exercise difficulty (Beinborn et al., 2016). Additionally, homonymous or polysemous words are typically harder to process and acquire than words with just one meaning (Bensoussan and Laufer, 1984). Psycholinguistic features, such as the age of acquisition (Kuperman et al., 2012; Dascalu et al., 2016) and concreteness of the word (Ellis and Beaton, 1993; Van Hell and Mahn, 1997; De Groot and Keijzer, 2000), are important factors for word difficulty prediction, too. The initial letter plays a disproportionate role in word recognition and production: speakers in tip-of-the-tongue states often remember the first letter and phonological primes aid retrieval (Brown and McNeill, 1966; James and Burke, 2000).

L1-specific difficulty. Beyond L2-intrinsic word properties, difficulty is shaped by the learner's first language. Learners draw on L1 knowledge when acquiring an L2 vocabulary, with positive transfer where the languages are similar and interference where surface similarity is misleading (Odlin, 1989). At the lexical level, cognateness is a well-studied transfer mechanism. Cognates reduce learning effort because learners can map the L2 target onto the well-known L1 form: De Groot and Keijzer (2000) show that cognateness predicts L2-vocabulary retention, and Beinborn et al. (2014) find that learners infer the meaning of unfamiliar words more readily when there is a cognate in their L1. Similarly, Urdaniz and Skoufaki (2022) find that frequency and cognateness are the two

strongest predictors of lexical difficulty for Spanish learners of English, with significant interaction such that word frequency is more predictive for non-cognates than cognates. However, coincidental similarity or shifted meaning of cognates can cause interference (Odlin, 1989; Bensoussan and Laufer, 1984).

These findings imply that the same word can be easy for one learner and hard for another, depending on the overlap of L1 and L2. Moreover, cognateness interacts with other predictors of difficulty, motivating models that jointly incorporate L1-specific and L2-intrinsic information.

2.2 Predicting lexical difficulty

Progress on predicting lexical difficulty[3] has been driven by shared-task competitions (Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021). Computational approaches to predicting lexical difficulty have shifted from binary classification of complex words to continuous regression of difficulty scores (North and Zampieri, 2023). Across studies, robust predictors include word frequency, word length, age of acquisition, and concreteness (North et al., 2023).

These patterns extend beyond English. The CWI-2018 shared task shows that models trained on English, German, and Spanish can generalize to French, suggesting that certain cognitive mechanisms transfer across typologically related languages (Yimam et al., 2018). The BEA 2024 shared task investigated lexical-complexity prediction and simplification with 10 languages, finding that feature-based systems can compete with large language models (Shardlow et al, 2024). For some languages, script-specific features are important, e.g., logographic-character frequency in Chinese and Japanese (Lee and Yeung, 2018a; Nishihara and Kajiwara, 2020). Furthermore, Finnimore et al. (2019) confirm features from typologically related languages can be beneficial, while including features from unrelated languages may reduce performance.

Despite these advances, learner-specific factors remain underexplored. Studies that model the L1 background report benefits over agnostic models, e.g., for Dutch learners of French as well as Chinese and Spanish learners of English (Tack et al., 2016; Lee and Yeung, 2018b; Urdaniz and Skoufaki, 2022). However, these studies focus on a single language pair. Palmero Aprosio et al. (2020) extend this line of research to multiple L1s for learners of Italian by modeling cognates and false friends. Similarly, the lexical complexity for learners of Japanese has been studied for several L1 backgrounds (Ide et al., 2023; Nohejl et al., 2024). Still, the interaction between L1-specific and general features remains largely unexamined. In this article, we address that gap directly.

3 Experimental setup

We train and evaluate L1-specific CatBoost models (Prokhorenkova et al., 2018) for predicting the difficulty of English vocabulary items. This section describes the dataset, feature groups, model, and evaluation metrics.

3.1 Data

We use the Knowledge-based Vocabulary Lists (KVL), a large-scale dataset of lexical difficulty based on vocabulary test responses from more than 100,000 second-language learners of English (Schmitt et al., 2021; Skidmore et al., 2025) as part of the BEA 2026 shared task (Felice and Skidmore, 2026). The data covers three L1s—Spanish, German, and Chinese—with 6,091 training, 677 development, and 748 test items per L1.

In the KVL test format, learners see a word in their L1 with a context sentence and must type the English translation, given the first letter and blanks for the remaining letters:

> *Kabel* German, noun
> *Achtung, stolpere nicht über das Kabel am Boden.*
> `c _ _ _ _`

Each item is scored as correct or incorrect. In the example above, the correct completion would be `able`. Responses are aggregated into a continuous logarithmic difficulty score by a Rasch model (De Boeck, 2008; Dunn, 2024), which jointly estimates item difficulty and learner ability such that the resulting centered score reflects the intrinsic difficulty of a word (higher scores are easier).[4]

The difficulty estimates may vary by L1, see Fig. 1. The word `cable` is easiest for Spanish speakers (score 2.3) because the corresponding

[3] In this article, we use *difficulty* to refer to the processing effort in retrieval, as defined in Bulté et al. (2025), and distinguish this from structural complexity, which concerns the internal composition of a word.

[4] Note that the values across languages are not directly comparable in an absolute way but rather as relative to the average word difficulty of an L1.

Spanish form is spelled identically; moderately easy for German speakers (score 0.9), who benefit from the German cognate *Kabel*; and hardest for Chinese speakers (score 0.0, 电缆 *diànlǎn*).

3.2 Features

We extract 24 features per item and organize them into four feature groups that structure our analysis, see Fig. 1. Each group targets a different aspect of what makes a word easy or difficult to recall. Full definitions and sources are provided in Appendix C. Here, we describe the rationale behind each group.

Familiarity (11 features). These features capture how likely a learner is to have encountered a word before. We include logarithmic word frequency and contextual diversity from SUBTLEX-UK (Van Heuven et al., 2014), reported age of acquisition and percentage of annotators who knew the word (Kuperman et al., 2012), CEFR level in the form of CEFR-J proficiency level, e.g., B2 for `cable`, and the EFLLex word-frequency profile across CEFR-graded books (Europarat, 2011; Negishi et al., 2013; Dürlich and François, 2018). Together, these features reflect both naturalistic exposure (how often the word appears in English media) and curricular sequencing (at which proficiency level the word is typically introduced).

Meaning (5 features). To approximate semantic complexity and concreteness, we use the ℓ_2 norm of the English fastText embedding (Bojanowski et al., 2017), mean hypernym depth and sense count from WordNet[5] (Fellbaum, 1998), the part-of-speech (POS) ratio from SUBTLEX-UK, and a binary flag indicating whether the KVL item required disambiguation, e.g., for the item `decode`, the German translation is *etw entschlüsseln (nicht: decipher)*.

Surface (7 features). Surface features describe the orthographic form of the item as presented in the test. They include target- and source-word length in characters, number of syllables and letters per phoneme (an orthographic transparency proxy) of the target word, context-sentence length, and the first letter of both the English clue and the L1 translation. These features are relevant because the KVL task requires learners to reconstruct a spelling, making word length, letter cues, and phoneme density directly task-relevant (James and Burke, 2000; Beinborn et al., 2016).

Transfer (1 feature). We compute the cosine similarity between character n-gram TF-IDF vectors of the English word and its L1 translation, see Appendix D for a detailed description. This feature captures orthographic overlap as a proxy for cognateness: High values indicate transparent cognates (e.g., English `fantasy` and Spanish *fantasía*), while the similarity is zero for all Chinese items due to the different script. Although this measure does not account for phonological similarity or regular sound correspondences, we find it to be highly informative for languages that share the Latin script with English.

3.3 Model and baselines

For our purposes, we require a fast and interpretable model. We use CatBoost (Prokhorenkova et al., 2018), a gradient-boosted decision-tree method that handles both numeric and categorical features. We train one model per L1 on the same set of 24 features (Fig. 1, right), allowing each model to learn L1-specific interactions between features and difficulty. Hyperparameters are given in Appendix E. Model training takes about ten seconds. To estimate variance, we train each model with 20 random seeds.

We compare against two baselines: a ridge regression model trained on the same features, which tests whether nonlinearity or feature interactions contribute beyond a linear combination, and the transformer-based approach of Skidmore et al. (2025), which fine-tunes a pretrained language model directly on KVL difficulty scores without handcrafted features.

3.4 Evaluation

We report the root-mean-square error (RMSE), in alignment with the BEA 2026 shared task (Felice and Skidmore, 2026), and Pearson's r on the test set. Our team's submission results (*Boosted Cats – HuDS lab*) are available at the shared-task repository. To test whether L1-specific models capture shared or distinct difficulty structures, we additionally evaluate each model on the test sets of the other two L1s.

For interpretability, we use tree-based SHAP values (Lundberg and Lee, 2017; Lundberg et al., 2018), which attribute each prediction to individual feature contributions through Shapley values from

[5] For example, the word `cable` has six senses, one of which has a hypernym path of depth 9: cable ⊆ conductor ⊆ device ⊆ instrumentality ⊆ artifact ⊆ whole ⊆ object ⊆ physical entity ⊆ entity.

cooperative game theory (Shapley, 1953). We aggregate absolute Shapley values by feature group to obtain their relative importance, analyzed in Section 4.2. We compute relative group importance as the sum of absolute Shapley values within each feature group, normalized to sum to one per item.

4 Results

The same English word can be easy for one L1 group and hard for another, illustrated by the gold-label lexical difficulties that correlate across L1s but are far from identical, see Appendix F. This observation motivates our approach to train L1-specific models, asking three questions. Are the model predictions accurate enough to warrant a study of how these predictions arise (Section 4.1)? Which mechanisms do the models rely on, and how do these mechanisms differ by L1 (Section 4.2)? And finally, do models trained on one L1 generalize to another (Section 4.3)?

4.1 Prediction quality

Table 1 compares our CatBoost models against two baselines: a linear (ridge) regression trained on the same features and the transformer-based approach of Skidmore et al. (2025). CatBoost achieves the lowest RMSE on all three L1s and the highest correlation in German and Chinese. The improvement over the linear baseline shows the importance of accounting for feature interactions and non-linear contributions to difficulty. CatBoost outperforms the transformer baseline, too, implying that well-chosen psycholinguistic features can perform on par with pretrained representations while being more easily interpretable.

	RMSE ↓			Pearson's r ↑		
Model	**ES**	**DE**	**CN**	**ES**	**DE**	**CN**
Transformer	1.26	1.26	1.14	**0.77**	0.77	0.75
Linear regression	1.30	1.20	1.07	0.72	0.74	0.77
CatBoost (ours)	**1.24**	**1.12**	**1.04**	0.76	**0.78**	**0.79**

Table 1: Test-set performance per L1. CatBoost results report the median over 20 random seeds; the 5-95% range is ≤ 0.02 for all entries. Best result per L1 in **bold**. The shared-task baseline fine-tunes a transformer on KVL difficulty scores (Skidmore et al., 2025).

Figure 2 shows predictions against the gold labels per L1. The predictions correlate well with the gold labels, as the bulk of items clusters near the diagonal (dashed gray). The hardest items, however, tend to be predicted as easier than they are. This bias is consistent across L1s, suggesting that the most difficult words involve challenges not captured entirely by our feature set, e.g., rare senses (`received` in the meaning of "generally accepted"). Other highlighted examples include `question`, which is easy and well-predicted for all L1s, although it is a cognate only in Spanish; `to picture` and `handbook` show that our model captures the L1-dependent difficulty.

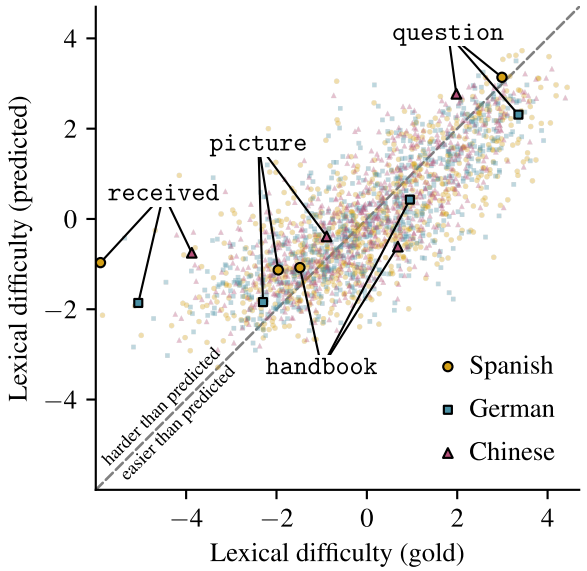

Figure 2: Predicted versus gold-label lexical difficulty. Each point is one L1-English vocabulary pair; the dashed diagonal marks a perfect prediction. Four exemplary items are highlighted: `question` (noun), `picture` (verb), `handbook` (noun), `received` (adjective). Seed variation of the prediction (5^{th}–95^{th} percentile) is small (median 0.1, max. < 0.4).

4.2 Feature-group importance

To understand the model predictions more intuitively, we thus group our 24 features into four feature groups—*familiarity*, *meaning*, *surface*, and *transfer*—and aggregate the absolute Shapley values per item for each group. The importance share of a feature group measures how strongly the features within this group contribute to predicting a specific item. We report the average importance of every feature and its correlation with the gold-label difficulty in Table 2.

Transfer differentiates Spanish and German from Chinese learners. The most prominent cross-linguistic difference is the role of transfer. For Spanish and German speakers, a single feature—character-n-gram cosine similarity between the L1 and the English word—is the

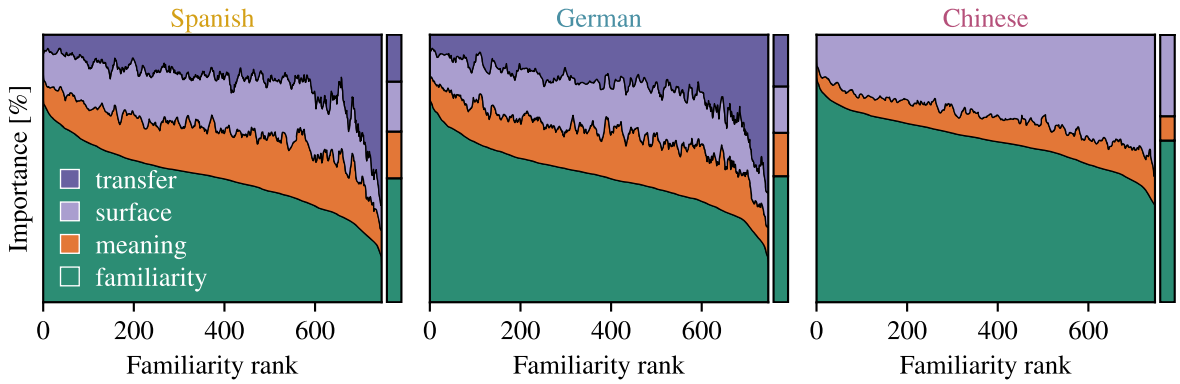

Figure 3: Per-item feature-group-importance shares, sorted by decreasing importance of familiarity (left to right). Median across 20 seeds, with a rolling average of window size 10. The bar at the right of each panel shows for each group the median across seeds of its mean importance share over all item. The *form* group is split into *transfer* (character similarity between the L1 and English) and *surface* (features like word length and initial letters).

strongest predictor overall (mean |SHAP| of 0.51 for Spanish and 0.52 for German; see Fig. 8 in Appendix C). For Chinese, this feature does not contribute at all because the Chinese writing system shares no characters with English.

This asymmetry is visible in Fig. 3, which shows the group importance of each item, sorted by decreasing importance of familiarity. Transfer fills the gap of less familiar words for Spanish and German. These are typically cognates, e.g., handbook and *Handbuch*, which a German speaker can infer even without prior exposure to the English word.[6] Transfer is most relevant for a distinct small subset of roughly 10–20% of items relying least on familiarity (above rank 600).

Familiarity is strong across L1s. Familiarity accounts for the largest average feature-group importance for all three L1s, see the stacked bar summaries in Fig. 3. Within this group, the most important features differ: age of acquisition (AoA) and frequency for Spanish and German, versus the proficiency levels (EFLLex-level span and CEFR-J levels) for Chinese. This difference may reflect a variation in how English is acquired—naturalistic exposure for Spanish and German learners as opposed to potentially more structured, curriculum-driven learning for Chinese learners. Yet in all cases, familiar words are easier to recall.

Surface and meaning. Surface features are as important as meaning features (embedding norm, hypernym depth, and sense count) for Spanish and German, and substantially more important for Chinese, see Fig. 3. The importance of surface features is likely due to the KVL task format, which requires form recall rather than correct usage of a word. Spelling difficulty makes surface cues, e.g., word length and priming by the clue letter, directly task-relevant (James and Burke, 2000; Beinborn et al., 2016). Chinese learners appear to rely on these cues more heavily.

Two routes to easiness. Figure 4 projects each item onto a triangle. Each corner represents that the respective feature group is uniquely important, while the middle corresponds to an equal mixture of importance. The color of each point encodes gold-label difficulty and the background shading is an estimate of the difficulty distribution in this triangle.[7]

For Spanish and German, easy items (green) cluster in the two lower corners, familiarity and form. These clusters represent two routes through which a word is easy to a learner. A word is easy because it is frequent and early-acquired or because it closely resembles its L1 form. Some words are both common and have high orthographic similarity, which makes them easy as well (see Appendix G for details), but they land in the middle region since the Shapley values for familiarity and transfer are both high. Difficult items (red) occupy the middle region in the triangle because neither cue is strong.

For Chinese, the distribution in the triangle is instead tightly clustered along the lower-left

[6]Transfer beyond orthographic overlap exists but is not captured by character similarity. For example, the Chinese word 手册 (*shǒucè*, literally "hand-book") elicits transfer, too.

[7]A Nadaraya–Watson kernel regression of the difficulty surface (Nadaraya, 1964; Watson, 1964) with bandwidth selected by leave-one-out cross-validation (Härdle, 1990).

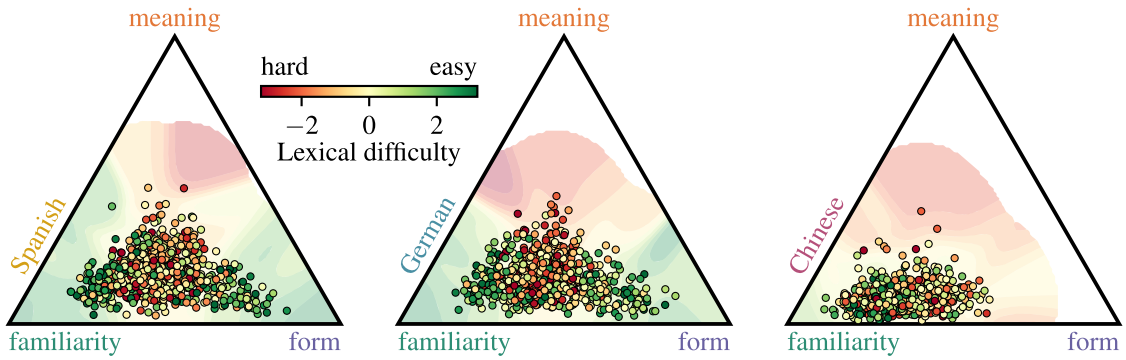

Figure 4: Each item projected onto a triangle according to the relative importance of three feature groups (familiarity, meaning, and form = surface ∪ transfer). Color encodes the gold-label difficulty of an item (red: hard, green: easy). Background shading shows a regression of the difficulty surface; regions with insufficient data are masked.

familiarity-form edge[8] and the background shading has no clear gradient. Chinese difficulty is shaped more uniformly by a combination of familiarity and form features.

4.3 Cross-L1 transfer of models

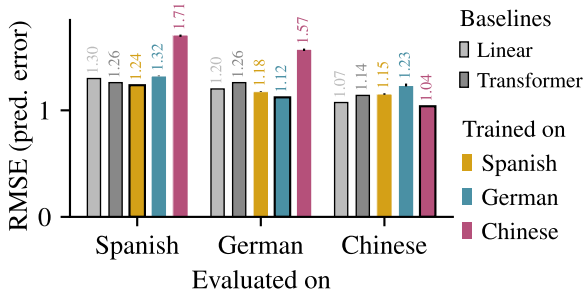

Figure 5: Cross-L1 evaluation. Each colored bar shows the RMSE prediction error (lower is better) of a CatBoost model trained on one L1 and evaluated on the L1 indicated on the horizontal axis. Thick black outlines mark within-L1 evaluation. Gray bars show the ridge linear baseline and the transformer baseline, each trained on the target L1. CatBoost results report the median over 20 random seeds with 5th–95th percentiles.

If Spanish and German learners rely on similar feature groups, a model trained on one L1 should reasonably generalize to the other. We test this by evaluating each L1-specific CatBoost model on all three test sets, see Fig. 5.

Each model performs best on its own L1, but the degradation when transferring between Spanish and German is minor. The Spanish-trained model achieves RMSE 1.18 on German (vs. 1.12 within-German), and the German-trained model reaches RMSE 1.32 on Spanish (vs. 1.24 within-Spanish)—both on par with the linear baselines (RMSE 1.30 for Spanish, 1.20 for German) and the transformer baselines (RMSE 1.26 for Spanish, 1.26 for German) trained directly on the target L1. A model trained on Spanish or German thus captures meaningful mechanisms that transfer to the other, without any access to target-L1 training data.

Transfer involving Chinese is markedly worse in both directions. Chinese-trained models yield RMSE 1.71 on Spanish and 1.57 on German, well above the baselines. Spanish- and German-trained models evaluated on Chinese (RMSE 1.15 and 1.23) degrade less severely, which aligns with the observation that the Spanish and German models have learned all the features relevant to predicting difficulty for Chinese learners, although they still fall short of the within-Chinese model (RMSE 1.04). The Chinese-trained model, not having learned to predict items based on transfer, cannot generalize to Spanish or German.

These cross-L1 results mirror the feature-group analysis: Spanish and German models share a difficulty landscape, while the prediction for Chinese learners relies on different cues. Practically, this observation suggests that models trained on a given L1, say Spanish, could provide useful estimates for typologically related languages such as Portuguese or Italian, even without the dedicated training data.

5 Discussion

Our finding that word familiarity exerts the strongest impact on the KVL difficulty scores across all L1 groups aligns with prior research identifying word frequency and age of acquisition as robust predictors of lexical processing (Rott, 1999;

[8] Chinese items show a somewhat lower dispersion in the feature-group simplex (Aitchison total variance: CN 0.30, ES 0.33, DE 0.35 with bootstrap 90% CIs < 0.05).

Ellis, 2002; Kuperman et al., 2012; North et al., 2023). Our analysis adds a quantitative decomposition that reveals *how* this shared component interacts with L1-specific factors. For Spanish and German, transfer provides a second route to easiness: Words can be easy because they are familiar or because they are orthographic cognates. This two-route structure aligns with Urdaniz and Skoufaki (2022), who find frequency and cognateness to be strong, interacting predictors for Spanish learners of English, and with De Groot and Keijzer (2000), who show that cognates are easier to remember. We observe the same pattern for German learners. For Chinese learners, difficulty is predicted more uniformly by surface features and familiarity. The cognitive strategies during form recall thus appear to differ qualitatively, depending on the typological relationship between the L1 and L2.

Our character similarity feature captures character overlap as a proxy for perceived similarity (Ringbom, 1987), which need not coincide with etymological relatedness. The strength of this feature in our models suggests that even coarse character overlap is a powerful retrieval cue when L1 and L2 share a script. Although Chinese has no orthographic overlap, cross-linguistic transfer may still be present in the form of loanwords (e.g., 雷达 *léidá* from English radar), calques (brainwash from the Chinese 洗脑 *xǐnǎo*), and parallel compositions (星光 *xīngguāng* happening to map to "starlight") or morphology (地 *de* marking adverbs like the English suffix "-ly").

The task format affects which aspects of word knowledge are tested (Laufer and Goldstein, 2004; Culligan, 2015). In our case, meaning features contribute little across all L1s, likely because the KVL task tests spelling rather than an understanding of meaning. A task requiring contextual usage would likely involve more meaning-related features.

A caveat concerns interpreting Shapley values as cognitively plausible. Our models approximate human test responses, and their strong held-out performance suggests that the features are plausible factors in the learner behavior that underlies the test responses. Where the model fails, however, the relevant factors likely lie outside our feature set, e.g., individual learner differences, contextual effects of the context sentence, or interference from false friends. We discuss the relation of lexical features beyond our predictor set with the gold-label difficulty and model-prediction errors in Appendix H.

Our findings also have practical implications. Spanish and German learners could benefit from curricula rich in cognates in order to efficiently build an extensive vocabulary (Nation, 2000). Furthermore, a model trained on data from one L1 could serve in predicting lexical difficulty for typologically related, lower-resource languages without dedicated data. For Chinese learners of English (and potentially speakers of languages not related to the target language), a curriculum based on frequency and proficiency levels is likely more effective. Since our model can generalize to unseen words, we provide predictions of lexical difficulty beyond the KVL dataset, on the entire SUBTLEX-UK vocabulary for which we could compute the features. Our model predictions can be explored interactively.

6 Conclusions

We have investigated what makes an English word difficult for second-language learners by training interpretable models on lexical-difficulty scores from Spanish, German, and Chinese L1 speakers. By decomposing the predictions into feature groups, we identify two qualitatively different profiles. Spanish and German learners benefit from two routes to easiness—familiarity and orthographic transfer—while Chinese learners rely on a less structured combination of familiarity and surface features in the absence of orthographic overlap. The models of learners with Spanish and German as L1 generalize well to each other, indicating that the learned difficulty functions reflect shared cognitive strategies.

For future research, the strong cross-L1 transfer between the Spanish- and German-trained models suggests that models trained on a high-resource L1 could provide useful difficulty estimates for typologically related languages (e.g., from Spanish to Portuguese, Catalan, or Italian; potentially also from Chinese to other languages not related to English like Japanese), even without the dedicated training data. Another promising direction is to investigate whether the identified mechanisms are symmetric, i.e., whether the same profile of feature-group importance applies for English speakers that learn Spanish, German, or Chinese. Third, incorporating phonological similarity and including calques and morphological similarity could capture part of the variance in lexical difficulty that our current feature set does not explain.

Limitations

Language coverage. The KVL data covers a limited selection of four high-resource languages, three of which are Indo-European. The extent to which the patterns we identify generalize to learners with other language backgrounds remains an open question.

Task specificity. The KVL dataset tests productive form recall only. Our analysis therefore pertains to this aspect of vocabulary knowledge and could be extended to other test formats.

Measuring transfer. We operationalize cross-linguistic transfer through character-level overlap, which does not capture phonological similarity, regular sound correspondences, and semantic transfer (e.g., calques).

Interpretation of feature importance. Feature-group importance describes the structure of model predictions and cannot measure a learner's cognitive processes directly. Inferences about human cognition thus remain indirect.

Curricular effects. Lexical difficulty is not only influenced by word properties but also the curriculum order, which is captured by our CEFR-proficiency-level features. Beyond the intrinsic difficulty of a word, the KVL difficulty we model therefore partially encodes the status quo in instruction regarding where a word happens to be placed in educational curricula.

Ethical considerations

This work is intended as fundamental research into L2 vocabulary difficulty. While our results may inform applications and further research, it should not be interpreted as prescriptive guidance for teaching in practice.

Our computational experiments use light-weight models that entail no substantial environmental impact.

Acknowledgments

We thank the reviewers for their thoughtful and detailed feedback. This research is partially supported by the zukunft.niedersachsen program of the VolkswagenStiftung (L.B., Z.H.) and by a VENI grant (Vl.Veni.211C.039) from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) (L.B., A.H.).

Author contributions

J.M.M.: Conceptualization, Analysis, Software, Literature review, Visualization, Writing—original draft, Writing—review & editing. **Z.H.:** Analysis and Software (feature analysis for App. H, KVL extension), Literature review, Visualization (App. H), Writing—original draft (App. B, D, and H; parts of Introduction and Related Work), Writing—review & editing. **A.H.:** Literature review, Writing—original draft (parts of Related Work). **L.B.:** Supervision, Conceptualization, Analysis, Literature review, Writing—review & editing.

References

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Readability for foreign language learning: The importance of cognates. *ITL - International Journal of Applied Linguistics*, 165(2):136–162.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2016. Predicting the spelling difficulty of words for language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 73–83, San Diego, CA, USA. Association for Computational Linguistics.

Marsha Bensoussan and Batia Laufer. 1984. Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, 7(1):15–32.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Roger Brown and David McNeill. 1966. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5(4):325–337.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2025. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*, 75(2):533–574.

Brent Culligan. 2015. A comparison of three test formats to assess word difficulty. *Language Testing*, 32(4):503–520.

Mihai Dascalu, Danielle McNamara, Scott Crossley, and Stefan Trausan-Matu. 2016. Age of exposure: A model of word learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, Phoenix, AZ, USA. AAAI Press.

Paul De Boeck. 2008. Random item IRT models. *Psychometrika*, 73(4):533–559.

Annette M. B. De Groot and Rineke Keijzer. 2000. What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, 50(1):1–56.

Karen J. Dunn. 2024. Random-item rasch models and explanatory extensions: A worked example using L2 vocabulary test item responses. *Research Methods in Applied Linguistics*, 3(3):100143.

Luise Dürlich and Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nick C. Ellis. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.

Nick C. Ellis and Alan Beaton. 1993. Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, 43(4):559–617.

Europarat, editor. 2011. *Common European framework of reference for languages: Learning, teaching, assessment*, 12th edition. Cambridge University Press, Cambridge, UK.

Mariano Felice and Lucy Skidmore. 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)*, San Diego, CA, USA. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*, 1st edition. The MIT Press, Cambridge, MA, USA.

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. Strong baselines for complex word identification across multiple languages. In *Proceedings of the 2019 Conference of the North*, pages 970–977, Minneapolis, MN, USA. Association for Computational Linguistics.

Wolfgang Härdle. 1990. *Applied Nonparametric Regression*, 1st edition. Cambridge University Press, Cambridge, UK.

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.

Lori E. James and Deborah M. Burke. 2000. Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6):1378–1391.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Batia Laufer and Zahava Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3):399–436.

John Lee and Chak Yan Yeung. 2018a. Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–4, Algiers. IEEE.

John Lee and Chak Yan Yeung. 2018b. Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, NM, USA. Association for Computational Linguistics.

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, Long Beach, CA, USA. Curran Associates Inc.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Èlizbar A. Nadaraya. 1964. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142.

Ian Stephen Paul Nation. 2000. *Learning Vocabulary in Another Language*, 1st edition. Cambridge University Press, Cambridge, UK.

Masashi Negishi, Tomoko Takada, and Yukio Tono. 2013. A progress report on the development of the CEFR-J. In Evelina D. Galaczi and Cyril J. Weir, editors, *Exploring language frameworks: Proceedings of the ALTE Kraków Conference, July 2011*, 1st edition, number 36 in Studies in language testing. Cambridge University Press, Cambridge.

Daiki Nishihara and Tomoyuki Kajiwara. 2020. Word complexity estimation for Japanese lexical simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3114–3120, Marseille, France. European Language Resources Association.

Adam Nohejl, Akio Hayakawa, Yusuke Ide, and Taro Watanabe. 2024. Difficult for whom? A study of Japanese lexical complexity. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 69–81, Miami, FL, 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*, 6:1236963.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*, 1st edition. Cambridge University Press, Cambridge, UK.

Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. 2023. Norm of word embedding encodes information gain. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2108–2130, Singapore. Association for Computational Linguistics.

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, CA, USA. 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*, pages 192–200, Genoa Italy. ACM.

Elke Peters. 2019. Factors affecting the learning of single-word items. In Stuart Webb, editor, *The Routledge Handbook of Vocabulary Studies*, 1st edition, Routledge handbooks in linguistics, pages 125–142. Routledge, Taylor & Francis Group, London, UK.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: Unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 6639–6649, Montréal, Canada. Curran Associates Inc.

Real Academia Española. 2025. Corpus del Español del Siglo XXI (CORPES).

Håkan Ringbom. 1987. *The Role of the First Language in Foreign Language Learning*, 1st edition. Number 34 in Multilingual matters. Multilingual Matters, Clevedon, UK.

Håkan Ringbom and Scott Jarvis. 2009. The importance of cross-linguistic similarity in foreign language learning. In Michael H. Long and Catherine J. Doughty, editors, *The Handbook of Language Teaching*, 1st edition. Wiley, Clevedon, UK.

Susanne Rott. 1999. The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, 21(4):589–619.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2021. Introducing knowledge-based vocabulary lists (KVL). *TESOL Journal*, 12(4):e622.

Norbert Schmitt and Diane Schmitt. 2020. *Vocabulary in Language Teaching*, 2nd edition. Cambridge University Press, Cambridge, UK.

Lloyd S. Shapley. 1953. A Value for n-Person Games. In Harold William Kuhn and Albert William Tucker, editors, *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press.

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.

Matthew Shardlow et al. 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.

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, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, Paris, France. AFCP - ATALA.

Raquel Perez Urdaniz and Sophia Skoufaki. 2022. Spanish L1 EFL learners' recognition knowledge of English academic vocabulary: The role of cognateness, word frequency and length. *Applied Linguistics Review*, 13(4):661–703.

Janet G. Van Hell and Andrea Candia Mahn. 1997. Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning*, 47(3):507–546.

Walter J. B. Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Geoffrey S Watson. 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372.

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, LA, USA. 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 Demo

We provide an interactive demo that allows users to explore the KVL dataset and our model predictions, see Fig. 6. Given an input text, words are highlighted according to their predicted or gold-label lexical difficulty. Selecting a word opens a panel showing its lexical difficulty, feature-group importance, top individual features, clue letter and context clue.

To map a word form in the input text to KVL lemmas, we precompute a lookup table of inflected forms using the PyInflect library. Because the demo is implemented as a static web application, it does not perform syntactic parsing on the fly. Instead, the most frequent part of speech is shown by default, with the option to select any alternatives. The interface also allows users to switch the language background and optionally include our extended vocabulary (see Appendix B).

B KVL extension

To support lexical-difficulty predictions beyond the original KVL inventory, we construct an extended vocabulary in the KVL task format derived from Wiktionary data.

Source and filtering. We begin with Wiktionary JSONL extracts produced by Wiktextract (Ylonen, 2022). From these, we retain only English headwords (`lang_code` = en) that appear in SUBTLEX-UK (Van Heuven et al., 2014). This filtering restricts the extension to commonly used words. We further discard entries that are empty, single-character, or all-uppercase (e.g., acronyms). Language-specific filters remove entries that primarily encode inflectional variants, spelling alternatives, or other metalinguistic notes.

Normalization. The retained entries are converted into a compact, standardized representation. From each Wiktionary entry, we extract only the information necessary for the KVL format: the English target word and part of speech, a masked spelling clue, an L1 translation, and an optional L1 context sentence. Duplicate English headwords are removed by keeping the first entry with a valid part-of-speech annotation.

L1-specific processing. For German and Spanish, the L1 source word is taken from the first available translation or, if unavailable, from the shortest gloss phrase. For Chinese, if no translation is available, the L1 source word is extracted from gloss blocks and normalized, including conversion to simplified characters. Context sentences are derived from example translations.

Excluding overlap. To ensure that the extended vocabulary contains no duplicates with the original KVL data, we exclude all English target words that occur in the KVL training, development, or test splits for the corresponding L1. The final output consists of three L1-specific CSV files that follow the original KVL schema and add 7,434 items for Chinese, 4,325 for German, and 6,606 for Spanish.

C Features

Table 2 lists the 24 features that we use for our models, grouped by mechanism. For each feature, we report the median absolute Shapley value per L1, computed by averaging absolute Shapley values over test-set items and then taking the median across 20 random seeds. We additionally report Spearman's rank correlation between the feature value and gold-label difficulty over test-set items.

Intuitively, the Shapley values quantify how important a feature is on average for the model predictions. Spearman's rank correlations show how monotonic the relation is between the feature value and lexical difficulty. Particularly interesting are those features for which the two measures do not align. For instance, character similarity has a weak correlation with difficulty for Spanish ($\rho = .10$) and German ($\rho = .25$), yet it is the feature with the highest Shapley-value importance for these two languages. This is in part because transparent cognates are relatively rare, and similarity is high only for this subset of transparent cognates. We discuss this in Appendix D. Moreover, the character similarity, as we show in our main analysis Section 4.2, interacts with familiarity such that this transfer feature tends to be important when familiarity is not available. Conversely, the CEFR-J level has the highest Spearman's $|\rho|$ across all L1s (up to $-.66$ for Chinese) but is only moderately important because it is collinear with other familiarity features, e.g., frequency and EFLLex-level span.

We briefly define features that are not self-explanatory:

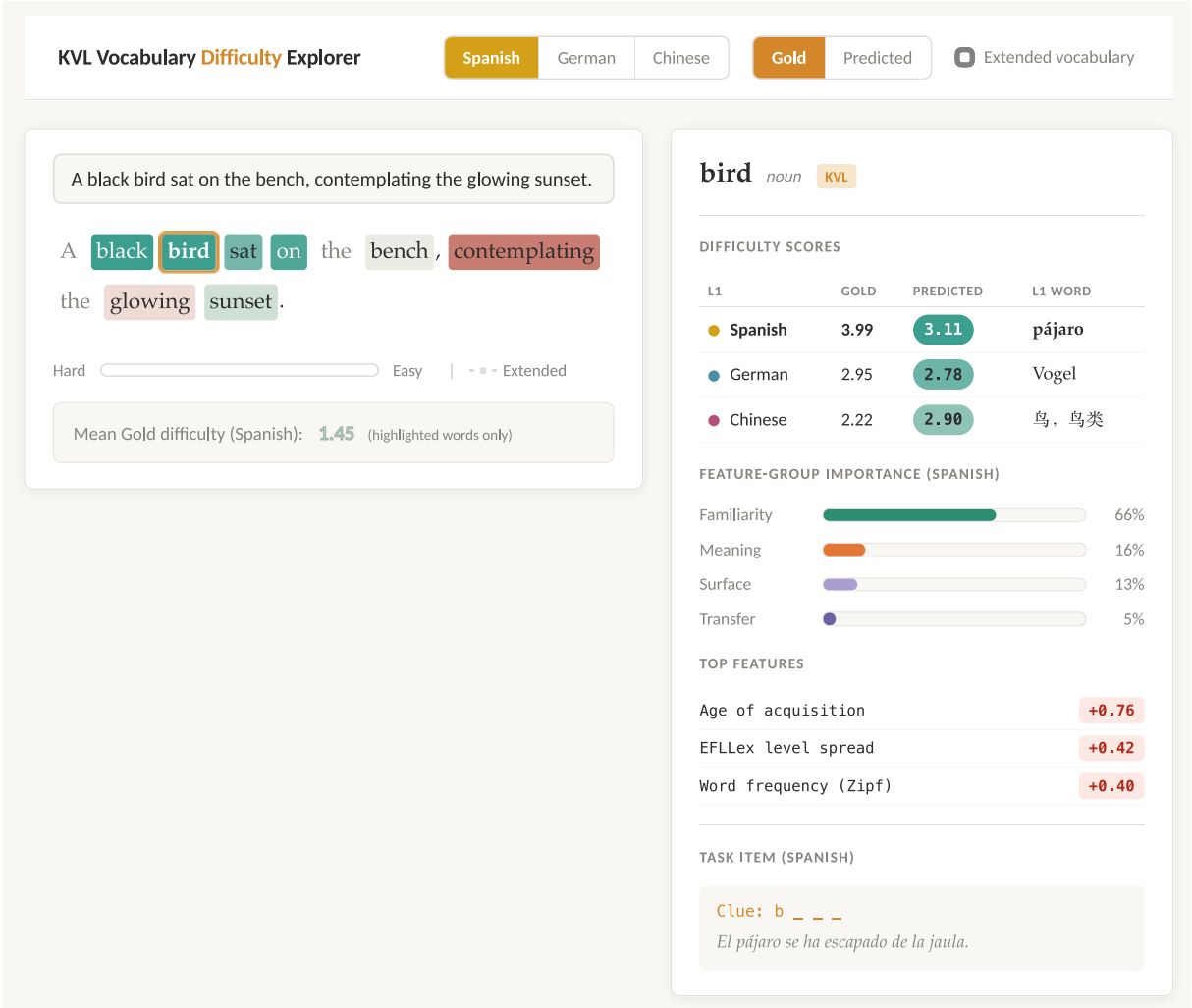

Figure 6: Screenshot of our interactive demo. Words of the input text are highlighted according to their lexical difficulty. Clicking on a word opens a panel that shows the gold-label and predicted difficulty as well as the feature-group importance. The L1-background can be switched.

Logarithmic frequency. Log-transformed word frequency on the Zipf scale (Van Heuven et al., 2014)

$$f_{\text{Zipf}} = \log_{10}(\text{fpmw}) + 3 \tag{1}$$

which modifies the standardized measure of frequency per million words (fpmw) by a logarithmic transformation and adding 3 to keep the measure positive even for low-frequency words, making it more intuitive. We assign a value below the minimal frequency threshold, $f_{\min} - 0.5$, to words missing from the SUBTLEX-UK list.

Contextual diversity. The proportion of film or television programs in SUBTLEX-UK containing at least one occurrence of the word, reflecting the breadth of contexts in which a learner might encounter it (Van Heuven et al., 2014). We expect that words that occur only in certain domains are less familiar (and thus more difficult to learn) relative to words of the same frequency that occur across many domains. This feature turns out to be especially relevant for Chinese-speaking learners.

EFLLex level span. The number of Common European Framework of Reference for languages (CEFR) levels (Europarat, 2011), A1, A2, B1, B2, C1, at which the word has non-zero frequency in the EFLLex learner corpus (Dürlich and François, 2018). A span of 5 means the word appears at every level from A1 to C1 and is thus broadly familiar, whereas a span of 1 means it is confined to a single proficiency band.

Embedding norm. The ℓ_2 norm of the 300-dimensional English fastText embedding $\mathbf{e}$ (Bojanowski et al., 2017),

$$\|\mathbf{e}\|_2 = \sqrt{\textstyle\sum_{i=1}^{300} e_i^2}\,. \tag{2}$$

	Feature	Definition	Source	Mean |SHAP| ES	Mean |SHAP| DE	Mean |SHAP| CN	ρ (value-gold) ES	ρ (value-gold) DE	ρ (value-gold) CN
Familiarity	Log frequency	Log-scaled word frequency	SUBTLEX-UK[1]	**0.26**	**0.20**	**0.19**	.40	.43	.60
	Contextual diversity	Distinct film contexts	SUBTLEX-UK[1]	0.05	0.06	**0.16**	.37	.39	.58
	Age of acquisition	Mean AoA rating	AoA[2]	**0.27**	**0.27**	**0.20**	−.44	−.46	−.55
	Percent known	% of raters knowing the word	AoA[2]	0.02	0.03	0.02	.17	.18	.22
	CEFR-J level	Vocabulary level (A1 = 1 … C2 = 6)	CEFR-J[3]	0.15	0.17	**0.23**	−.55	−.57	−.66
	EFLLex-level span	CEFR levels with non-zero frequency	EFLLex[4]	**0.25**	**0.22**	**0.31**	.43	.45	.64
	EFLLex A1	Learner-corpus token share at A1	EFLLex[4]	0.06	0.04	0.02	.40	.42	.49
	EFLLex A2	” at A2	”	0.13	0.10	0.05	.40	.48	.52
	EFLLex B1	” at B1	”	0.05	0.06	0.03	.30	.36	.46
	EFLLex B2	” at B2	”	0.03	0.05	0.08	.16	.18	.34
	EFLLex C1	” at C1	”	0.03	0.03	0.03	.05	.05	.18
Meaning	Embedding norm	fastText embedding ℓ_2 norm	fastText[5]	**0.23**	**0.18**	0.08	.32	.33	.33
	Hypernym depth	Mean synset depth in hypernym tree	WordNet[6]	0.13	0.11	0.03	.15	.18	.05
	Sense count	Number of synsets for word + POS	WordNet[6]	0.07	0.04	0.03	.12	.14	.24
	POS dominance ratio	Token share of dominant POS	SUBTLEX-UK[1]	0.04	0.03	0.05	.00	−.01	−.07
	L1 confusor flag	L1 source had disambiguation annotation	KVL[7] (derived)	0.02	0.05	0.01	−.08	−.12	−.09
Surface	Target word length	Character length of English word	KVL[7] (derived)	0.04	0.04	0.10	−.33	−.35	−.43
	Source word length	Character length of L1 lemma	KVL[7] (derived)	0.04	0.05	0.15	−.23	−.36	−.33
	Syllable count	Estimated syllables of English word	KVL[7] (derived)	0.05	0.05	0.05	−.27	−.32	−.37
	Letters per phoneme	Orthographic transparency proxy	AoA[2]	0.04	0.04	0.02	.08	.01	.10
	Context sent. length	Character length of L1 context sentence	KVL[7] (derived)	0.10	0.07	0.11	−.21	−.24	−.30
	Clue letter	First letter of the spelling clue	KVL[7]	0.18	0.12	0.16	—	—	—
	L1 initial letter	First letter of the L1 lemma	KVL[7] (derived)	0.06	0.07	0.07	—	—	—
Transfer	Character similarity	Char. n-gram TF-IDF cosine, EN vs. L1	KVL[7] (derived)	**0.51**	**0.52**	0.00	.10	.25	—

[1] Van Heuven et al. (2014); [2] Kuperman et al. (2012); [3] Negishi et al. (2013); [4] Dürlich and François (2018); [5] Bojanowski et al. (2017); [6] Miller (1995); Fellbaum (1998); [7] Schmitt et al. (2021); Skidmore et al. (2025)

Table 2: Mean |SHAP| importance per feature and L1, averaged over test-set items across 20 random seeds (5–95% range < 0.02 for all features and L1s), and Spearman's ρ between feature value and gold lexical difficulty. Features are grouped by category. The five highest |SHAP| values per L1 column are set in **bold**.

Since function words and polysemous words tend to have smaller norms because their embeddings are pulled towards the origin of the space by diverse contexts, concrete and semantically specific words tend to have larger norms (Oyama et al., 2023).

Character similarity. We represent the L1 word and its English translation as character n-gram TF-IDF vectors ($n \in \{2, 3, 4\}$, sublinear TF) and compute their cosine similarity. For Chinese, all values are approximately zero because there is no character overlap between Latin and Chinese script. We give a more detailed definition in Appendix D.

POS dominance ratio. The proportion of SUBTLEX-UK tokens for a lemma that carry its most frequent part of speech (POS), derived from the POS frequency in SUBTLEX-UK (Van Heuven et al., 2014). Values near 1 indicate unambiguous POS. Lower values indicate that the word is frequently used as multiple parts of speech (e.g., `light` as noun, verb, and adjective).

Letters per phoneme. The ratio of character length to phoneme count, derived from the number of phonemes per word in the Kuperman et al. (2012) dataset. This serves as a proxy for orthographic transparency: Words with many letters per phoneme have less predictable spellings, e.g., `thought` /θɔːt/ (7 letters / 3 phonemes = 2.33).

Clue letter and L1 initial letter. These categorical features encode the first letter of the English spelling clue and the first letter of the L1 lemma, respectively. They capture letter-level priming effects: certain initial letters may be associated with easier or harder vocabulary on average, and letter overlap between the clue and the L1 form may facilitate recall. Spearman correlations are not reported for categorical features. Interestingly, the L1-first-letter priming is equally important for Chinese as for Spanish and German (0.06–0.07) as the model learns a baseline for every Chinese word-initial character.

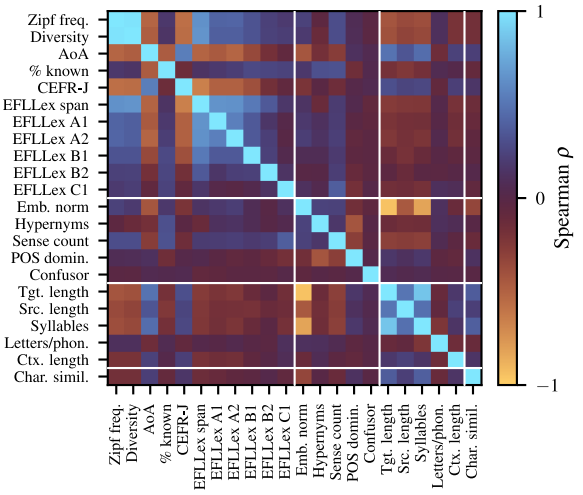

Figure 7: Pairwise Spearman correlation between numeric features (Spanish; the pattern is similar for German and Chinese except for L1-dependent features such as source word length and character similarity). Several pairs are substantially correlated. White lines indicate feature groups.

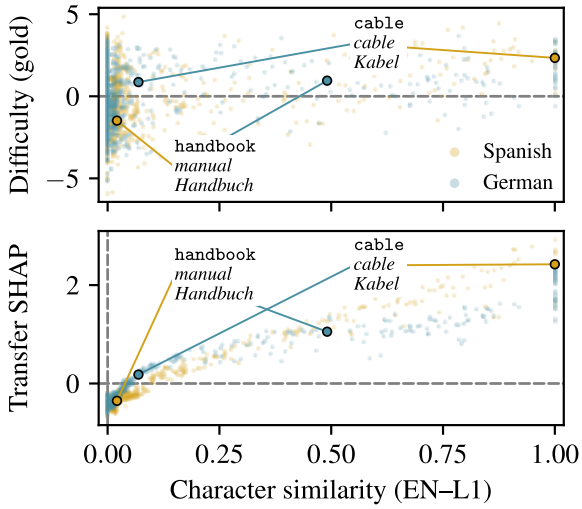

Figure 8: Character similarity between English words and their L1 translations (Spanish, German) versus gold-label difficulty (top) and Shapley-value contribution of this feature (bottom). Two words are highlighted: `cable` and handbook.

Figure 7 shows pairwise correlations between the numeric features from Table 2. Some features are strongly correlated, particularly the familiarity features (e.g., logarithmic frequency, contextual diversity, and EFLLex level span) and the surface features (target word length, syllable count, and source word length). This collinearity means that individual SHAP values within a group partly reflect shared variance, which motivates our decision to group the features for the main analysis.

Additional features were evaluated during development but excluded from the final model after unstructured ablation showed no consistent improvement. These include: character-level surprisal, normalized Levenshtein edit distance, longest common subsequence ratios, EFLLex entropy and mean level, WordNet synonym count and derivational family size, POS competition flag, and Spanish-specific CORPES frequencies and anglicism indicators. We provide these in Table 3 for completeness. These features turn out to be too redundant to yield higher performance. For example, edit distance captures the same signal as character similarity but is less informative in this context. We give an exemplary analysis in Appendix H.

D Character similarity

To approximate cross-linguistic form similarity between an English word and its L1 translation, we compare their character-n-gram overlap. This feature is motivated by the intuition that words sharing many letter sequences (e.g., `fantasy` and *fantasía*) can be readily recognized as cognates.

From binary overlap to weighted similarity. In a naive approach, we could represent each word as a binary vector over character n-grams, where each entry states whether a given n-gram occurs in the word, and then compute the cosine similarity between two vectors. This approach would treat all n-grams the same, yet some n-grams might be more informative than others.

To account for the informativeness of an n-gram, we weight it using *term-frequency–inverse-document-frequency* (TF-IDF). In this context, words correspond to "documents" and n-grams with $n \in \{2, 3, 4\}$ to "terms". The TF-IDF weight of an n-gram t in a word w is implemented as

$$\text{tf-idf}(t, w) = \big(1 + \log(\text{tf}(t, w))\big) \log\left(\frac{N}{\text{df}(t)}\right), \tag{3}$$

where $\text{tf}(t, w)$ is the frequency of t in w, $\text{df}(t)$ is the number of words in the training data containing t, and N is the total number of word forms in the union of English and L1 word forms in the training data. We use sublinear term-frequency scaling, i.e., $1+\log(\text{tf}(t, w))$ instead of $\text{tf}(t, w)$, so that repeated occurrences are weighted less than linearly.

Cosine similarity. We can thus represent each word as a TF-IDF-weighted vector over character

Feature	Definition
Character surprisal	Mean $-\log(p(\text{char}))$ under training-set unigram distribution
Edit distance (norm.)	Levenshtein distance between EN and L1 word, normalized by $\sqrt{\text{len}_{\text{EN}}} \cdot \sqrt{\text{len}_{\text{L1}}}$
LCS ratio (EN)	Longest common subsequence length / target word length
LCS ratio (L1)	Longest common subsequence length / source word length
EFLLex entropy	Normalized Shannon entropy of the CEFR-level frequency distribution
EFLLex mean level	Frequency-weighted mean CEFR level
WN synonym count	Number of unique lemma names across all synsets
WN derivational family	Number of derivationally related forms in WordNet
WN POS count	Number of distinct parts of speech in WordNet
POS competition	1 if the tested POS differs from the dominant SUBTLEX-UK POS
Exact match	1 if EN word = L1 lemma after normalization
L1 second letter	Second letter of the L1 lemma
CORPES frequency (ES)	Log-transformed normalized frequency in the Spanish CORPES corpus
CORPES dispersion (ES)	Log-transformed dispersion across CORPES subcorpora
Anglicism frequency (ES)	CORPES frequency of the English word looked up in Spanish text
Anglicism flag (ES)	1 if the L1 lemma appears in a Wiktionary-derived anglicism list

Table 3: Additional features evaluated on the development set but excluded from the final model after ablation showed no consistent improvement of model performance. CORPES refers to Real Academia Española (2025).

n-grams. Cross-linguistic form similarity is then measured as the cosine similarity between the English word w_{en} and its L1 translation w_{L1}:

$$\text{sim}(w_{\text{en}}, w_{\text{L1}}) = \frac{\mathbf{x}_{w_{\text{en}}} \cdot \mathbf{x}_{w_{\text{L1}}}}{\|\mathbf{x}_{w_{\text{en}}}\| \|\mathbf{x}_{w_{\text{L1}}}\|}. \quad (4)$$

The TF–IDF vectors are ℓ_2-normalized, so cosine similarity corresponds to their dot product.

Empirical behavior. Figure 8 illustrates how this character similarity relates to lexical difficulty and to its feature importance in Spanish- and German-L1 model predictions. Most items have near-zero similarity, with some items spread across the range up to perfect matches at 1. Higher similarity is associated with lower difficulty, cf. Table 2. The Shapley values show that high similarity contributes positively to predictions above a small threshold at approximately 0.05–0.10. Below this threshold, negligible similarity is a cue for higher difficulty. This nonlinear distribution explains why character similarity has high average feature importance despite only a moderate correlation with difficulty.

E Hyperparameters

Relevant hyperparameters are summarized in Table 4. The hyperparameters were tuned on the Spanish development set. Reported results are based on 20 random seeds.

F Correlation of difficulties across L1s

Figure 9 shows pairwise correlations of gold-label difficulty across L1s. All language pairs are positively correlated, quantifying the shared component of lexical difficulty: Words that are hard for one L1 group tend to be hard for others. However, the correlations are far from perfect, indicating that L1-specific factors account for a substantial portion of the variance.

Component	Parameter	Value
CatBoost	Loss function	RMSE
	Tree depth	7
	Learning rate	0.017
	Iterations	2,400
	ℓ_2 leaf regularization	0.8
Evaluation	Eval. seeds	20 (1, 8, 15, …, 134)
	Submission seeds	3 (42, 142, 242)
	Metrics	RMSE (primary)
		Pearson's r
Baseline	Model	Ridge regression
	ℓ_2 regularization	1

Table 4: CatBoost model configuration, evaluation setup, and baseline.

G Relationship of frequency and transfer

In Fig. 10, panel (a) shows lexical difficulty in the plane of character similarity and word frequency. We observe that the easiest words are frequent or cognate. Those words that are both cognate and frequent (items in the top right) shape the model predictions equally through both features, as the Shapley values in panels (b) and (c) indicate. The thresholds for the Shapley values that the model learns appear largely independent for these two features: a vertical divide.

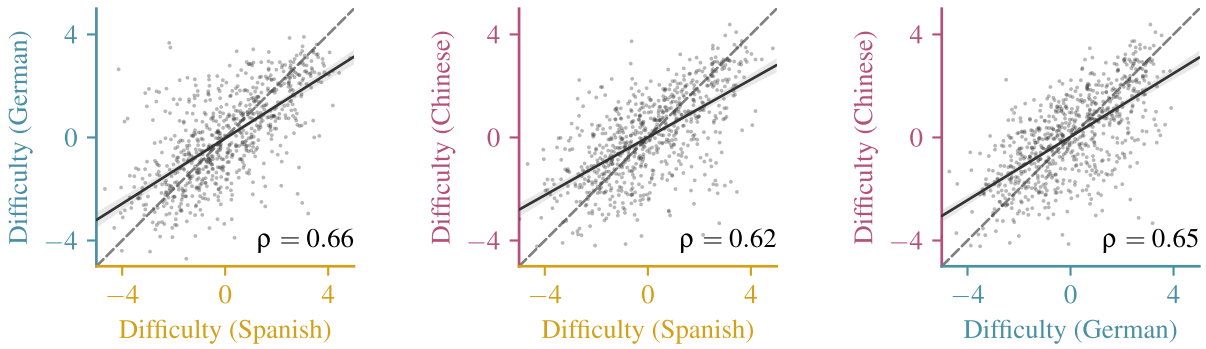

Figure 9: Pairwise correlation of gold-label lexical difficulty across L1s. Each point is one English word tested in both languages. Pearson's r is shown per pair. The shared variance reflects universal difficulty factors and the unexplained variance motivates L1-specific modeling.

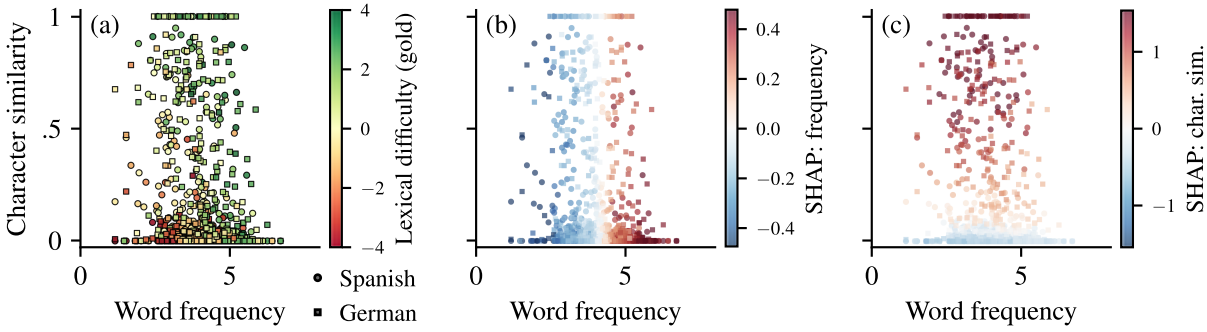

Figure 10: Character similarity by word frequency for Spanish and German, colored by (a) lexical difficulty and the feature importances of (b) word frequency and (c) character similarity.

H Feature analysis

In this section, we examine several features that we evaluated during development but excluded from the final model, cf. Table 3. The goal is to show how these features are related to lexical difficulty and why they may be redundant.

POS competition. Some English word forms occur frequently with one part of speech while another is less common, e.g., `received` as an adjective meaning "accepted" or `bar` as verb rather than a noun. We define a binary POS-competition flag, i.e., whether the POS of the target word differs from the most frequent POS.

We test whether POS competition affects the prediction error by comparing the prediction errors (gold − predicted) for items with and without POS competition, see Fig. 11. We find, across all L1s, no significant difference in mean or central tendency (e.g., Spanish: Welch's t-test: $p = 0.09$, Wilcoxon–Mann–Whitney test: $p = 0.36$), indicating that POS competition does not introduce a systematic model-prediction bias. Furthermore, items with POS competition do not differ significantly in gold-label difficulty from items without competition (Welch's t-tests and Wilcoxon–Mann–Whitney tests, all $p > 0.20$).

However, POS competition has a robust effect on the dispersion of the prediction error. Items with POS competition exhibit substantially higher variance across all L1s, with standard deviations increasing by 24-32%, see Table 5. This difference is statistically significant according to both Brown–Forsythe and Fligner–Killeen tests ($p < 0.01$), which suggests that POS competition is more difficult to predict for the model, but does not introduce a systematic bias.

L1	std_{no}	std_{yes}	ratio	BF p	FK p
Spanish	1.16	1.45	1.24	< 0.01	< 0.01
German	1.04	1.37	1.32	< 0.01	< 0.01
Chinese	0.97	1.24	1.28	< 0.01	< 0.01

Table 5: Standard deviation of the prediction error for items with and without POS competition across L1s, with their ratio, and p-values from Brown–Forsythe (BF) and Fligner–Killeen (FK) tests.

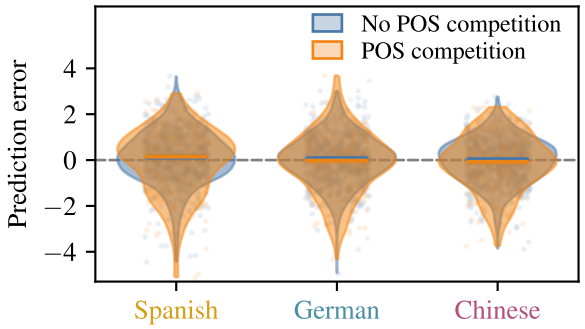

Figure 11: Prediction-error distribution (gold − predicted) by POS competition status across L1 groups (Spanish, German, Chinese). The violin plots compare items without POS competition ($n_{\text{no}} = 585$) versus items with POS competition ($n_{\text{yes}} = 163$) per language, with horizontal lines marking the median.

Edit distance. In addition to character-level cosine similarity, we evaluated character edit distance as a measure of orthographic similarity. While edit distance shows a weak-to-moderate negative correlation with gold-label difficulty for Spanish ($r = -0.17$) and German ($r = -0.30$), it is strongly negatively correlated with character-level cosine similarity for Spanish ($r = -0.77$) and German ($r = -0.73$), see Fig. 12. These two features encode largely the same information, making the inclusion of both redundant. Compared with cosine similarity, edit distance spreads out dissimilar word pairs over a wider range while compressing highly similar forms, making it less informative for distinguishing degrees of cognateness.

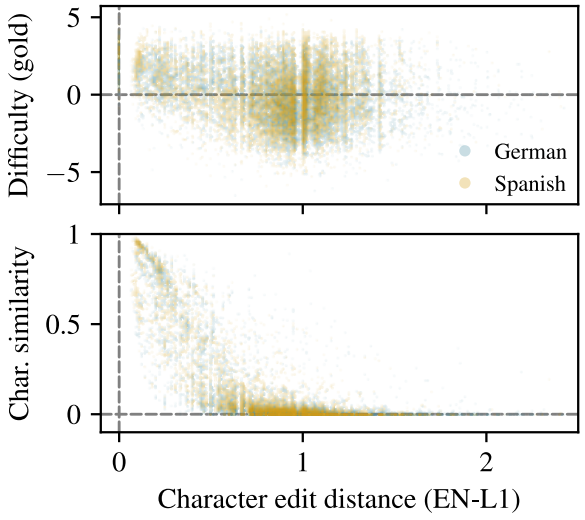

Figure 12: Normalized edit distance between English words and their L1 translations in Spanish and German versus gold-label difficulty (top) and its relationship with character-level cosine similarity (bottom).

L1 frequency. Finally, we examine L1 word frequency, which has also been suggested as a predictor of L2 lexical difficulty, under the assumption that more frequent words in a learner's native language may facilitate recognition or acquisition (Ellis, 2002; Peters, 2019). Using Spanish as an example, Fig. 13 shows that, although higher L1 frequency is associated with lower difficulty ($r = 0.15$), L1 and L2 frequency are themselves strongly positively correlated ($r = 0.44$). As a result, L2 frequency is largely redundant when L1 frequency is already included.

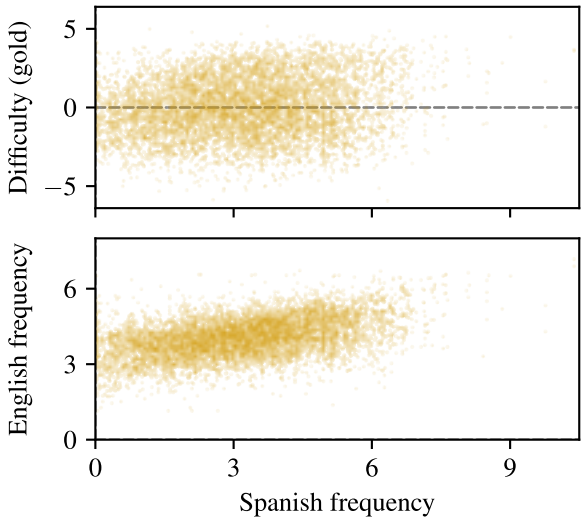

Figure 13: Frequency of L1 (Spanish) source words versus gold-label difficulty (top) and L2 (English) word frequency (bottom).