Translation Surprisal Predicts Learner Error in L2 French Production: A Language-Pair-Specific Difficulty Feature for ICALL

Matteo Lanza
King's College School
Southside, Wimbledon Common, London SW19 4TT, United Kingdom
matteo.lanza@icloud.com

Abstract

Word-difficulty measures in Computer-Assisted Language Learning (CALL) typically describe the target word itself, regardless of the learner's L1. Psycholinguistic research shows that translation probability predicts production, but evidence on learners comes largely from small laboratory studies using elicited norms, which are unavailable for English-French and do not distinguish the inflected form a learner must produce. From word alignments in a subtitle parallel corpus, we estimate the probability that each English word is rendered as each French form and score each learner production by the surprisal of the required form (how unexpected it is, given the English word). A mixed logistic model fitted to 384,047 typed productions from 1,213 learners in Duolingo's SLAM dataset shows that surprisal predicts error: a one-SD increase raises error probability from 21% to 30% and explains 24% of residual between-word variance. The effect lies in the probability of the specific form rather than the word's average uncertainty (entropy) or number of alternatives. Both lexical and inflectional choice contribute, and reverse-direction uncertainty does not increase error. Adding surprisal also doubles the correlation with held-out word-form difficulty. The measure is an interpretable, language-pair-specific difficulty feature for intelligent CALL (ICALL) sequencing and feedback, computable without learner data.

1 Introduction

Systems that predict learner error from platform logs, such as those developed for the Duolingo SLAM shared task, primarily use learner history and large feature sets (Settles et al., 2018); the few language-pair-specific features were form-similarity proxies for cognate status, such as edit distance between a word and its machine translation (Rich et al., 2018; Tomoschuk and Lovelett, 2018). Word-level difficulty in CALL has instead typically been predicted from properties of the target word, including frequency, length, age of acquisition, and CEFR level (Alfter and Volodina, 2018; Strohmaier and Buttery, 2024). Where language-pair-specific features have been used, they measure orthographic or semantic similarity between a word and its L1 translation (Hopman et al., 2018; Felice and Skidmore, 2026; Mayer Martins et al., 2026), or derive item difficulty from alignment-based translation distributions without testing it against learner error (Graën and Schneider, 2017; Chaudhary et al., 2021). To our knowledge, the probability of the translation a learner must produce has not been tested as a difficulty feature against learner production error.

Psycholinguistic research provides a potential language-pair-specific measure. Words with multiple translations are harder to learn and produce (Tokowicz and Kroll, 2007; Degani and Tokowicz, 2010) and slower to recognise, including for French-English bilinguals (Laxén and Lavaur, 2010), and the probability of the specific translation is a stronger predictor of performance than the number of alternatives (Prior et al., 2013). However, this evidence comes from laboratory studies with relatively small numbers of participants and items, using translation probabilities elicited for isolated words. Such norms are available for several language pairs, including English-Spanish (Prior et al., 2007), English-Dutch (Tokowicz et al., 2002), English-Chinese (Wen and van Heuven, 2017), and English-German and English-Hebrew (Degani et al., 2016), but not English-French. Moreover, isolated-word norms cannot capture which inflected form a learner must produce in context. Prior et al. (2011) found moderate convergence between elicited norms and corpus-derived translation probabilities, and alignment-

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 256–268, November 19–20, 2026, University of Tartu Library

based entropy correlates with elicited norms across language pairs (Lim et al., 2023), suggesting that corpus distributions can capture part of the same information. For translators, surprisal of the chosen rendering (how unexpected it is, given the source sentence) predicts production time (Lim et al., 2024); learners have not been tested.

This paper tests whether a corpus-derived translation probability predicts learner production at scale. We estimate, for each English word, the probability that it is rendered as each French form using word alignments in an English-French subtitle corpus (Lison and Tiedemann, 2016), and score each learner production by the surprisal of the required French form. For example, *know* is usually rendered by *savoir*: *sais* has a surprisal of 0.98 bits, whereas *connais* (from *connaître*) has 4.00 (Figure 2b). We test the measure on 384,047 typed productions from 1,213 learners in Duolingo's SLAM dataset (Settles et al., 2018), using a mixed logistic model with baselines for learner, word form, and exercise.

We ask:

1. Does the probability of the French form an English speaker must produce predict production error?
2. Is the effect driven by the probability of the specific form, the word's average uncertainty, or simply the number of alternatives?
3. Do lexical and inflectional surprisal contribute independently?
4. Does surprisal remain predictive after accounting for alternative lexical properties, practice, and direction of translation?
5. Can surprisal predict the difficulty of word forms that no learner in the training data has previously attempted?

2 Methods

2.1 Overview

Testing whether the probability of a required French form predicts learner error requires an independent estimate of that probability, a method for linking each French token to its English prompt word, and a model that separates this effect from learner ability, word difficulty, and exercise difficulty (Figure 1).

The outcome data are per-token correctness labels from Duolingo's SLAM 2018 dataset (Settles et al., 2018). The probability estimates come from an unrelated English-French subtitle corpus. Word alignments produce a distribution of French forms for each English word. Each Duolingo token is then linked to its English prompt word and scored by the surprisal of the required French form. Logistic mixed models with crossed random intercepts for learner, word form, and exercise test the resulting predictors.

2.2 Datasets

The learner corpus is the SLAM 2018 dataset (Settles et al., 2018), released by Duolingo for a second-language acquisition modeling shared task. It contains learners' first 30 days across English-French, English-Spanish, and Spanish-English tracks. We use the French-for-English-speakers training split only, with no learner-level filtering. Each exercise records the English prompt, learner ID, exercise format, and correctness label for each reference token. For example, if the prompt is *I know the girl* and a learner types *Je sais la fille*, the closest accepted answer is *Je connais la fille*, and only its token *connais* is labelled an error. The SLAM exercises consist of translating or transcribing short sentences or phrases, and each learner's exercises are recorded by the SLAM dataset, tagged by session type (lessons introducing new material, practice and tests), so the same words recur across sessions.

There are three exercise formats: (i) `reverse_translate` where a learner is given an English sentence and types its French translation; (ii) `reverse_tap` where a learner is shown an English sentence and constructs the French translation by tapping word tiles from a bank that contains the correct words but also distractors; and (iii) `listen` where a learner hears a French sentence and types what they hear.

The primary analysis uses `reverse_translate`, as it requires the learner to produce the form from memory, so the forms competing for production are those in the learner's own lexicon, which the Course Set (Section 2.5) approximates. `reverse_tap` is used only as a secondary replication because its available tile set is not recorded; `listen` is excluded because it provides no English prompt. Labels are based on the full accepted-answer set rather than a single reference answer.

The primary frame (analysis set) contains 384,047 tokens from 1,213 learners, 1,685 French word forms, 1,467 English prompt words, and 117,745 exercises, with a global error rate of 0.206.

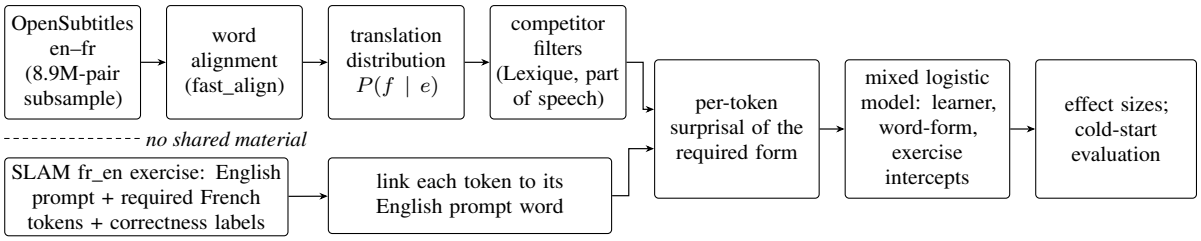


Figure 1: Method overview.

The parallel corpus is OpenSubtitles v2018 en-fr (Lison and Tiedemann, 2016; Lison et al., 2018), containing 41.8M aligned sentence pairs. The corpus is unrelated to the Duolingo material, so no sentence, learner or correctness label contributes to both the predictor and the outcome.

2.3 Translation distribution

For an English word e, the translation distribution is $P(f \mid e)$, the probability that e is rendered as French form f. We estimate this distribution from word alignments in the parallel corpus. Words are aligned with fast_align (Dyer et al., 2013) on a random 8.9M-pair subsample (Appendix A). Counts of English-French links are normalised by the English word's total links to obtain $P(f \mid e)$. The reverse distribution $P(e \mid f)$ is also calculated for token linking and reverse entropy.

2.4 Linking tokens to prompt words

The SLAM data provide an English prompt sentence but no direct word-level links. For each French token, every English word in the prompt is therefore scored using the product of forward and reverse translation probabilities and a positional distance penalty, $\exp(-4|\Delta \text{ normalised position}|)$. The highest-scoring English word is selected as the counterpart. Tokens with no scorable candidate (0.69% of translation-exercise tokens) are dropped.

Possessives and idioms can produce incorrect links and artificially high surprisal. A total of 372 tokens (0.1%) exceed 12 bits, mostly for this reason, and are excluded in a sensitivity analysis.

2.5 Competitor sets

An English word's competitors are the French forms it is aligned to. The distribution ignores sentence context: *le*, *la* and *les* are not free choices for *the* in a given sentence, so surprisal measures how dominant the required form is out of context. The raw alignment table contains substantial noise: the English word *the*, for example, has 12,275 competitors (the majority of which are a product of alignment noise across 9 million sentence pairs), compared with 586 for the average word. We therefore apply two filters. A French candidate must occur in Lexique (New et al., 2004), and its part of speech must match that of the English word (Appendix A). Elided forms are restored before dictionary lookup, and contracted articles (*au*, *du*, *aux*, *des*) are mapped to the lemma *le*, as for UD determiners (including indefinite *des*).

These filters reduce the average competitor set from 586 to 176 forms and that for *the* from 12,275 to 64, and translation probabilities are renormalised over the cleaned set. Most of these 64 forms are very improbable and carry little probability mass, so these low-probability tails have little effect on entropy or surprisal. Perplexity (2^H, the equivalent number of equally likely forms) is likewise barely affected by these tails: for *the* it is 6.4 (Figure 2a), about as uncertain as choosing among six equally likely forms.

The required French form is always retained; without this exemption, 24,351 tokens (6.3%) would receive infinite surprisal values.

The default competitor set contains all forms surviving these filters. A second "Course Set" restricts competitors to words in Duolingo's French vocabulary. Median coverage of the Course Set is 0.83 of each English word's translation probability (IQR 0.66–0.93). For competitor count, a French form must be linked at least five times, a floor fixed before any label was examined (Appendix A).

2.6 Entropy and surprisal

Entropy (Shannon, 1948) measures how evenly an English word's translation probability is distributed across French forms:

$$H(F \mid e) = -\sum_{f} P(f \mid e) \log_2 P(f \mid e). \quad (1)$$

A Miller-Madow small-sample bias correction is applied (Miller, 1955). Because French production

requires choosing both a lemma and an inflected form, entropy is decomposed as:

$$H(F \mid e) = H(L \mid e) + H(F \mid L, e), \quad (2)$$

where L is the lemma.

Surprisal instead describes the required token. If f^* is the required French form and l^* its lemma,

$$S = -\log_2 P(f^* \mid e). \quad (3)$$

Lexical and inflectional surprisal are defined as:

$$S_{\text{lex}} = -\log_2 P(l^* \mid e) \quad (4)$$

and

$$S_{\text{infl}} = -\log_2 P(f^* \mid l^*, e), \quad (5)$$

so that $S_{\text{lex}} + S_{\text{infl}} = S$ (Figure 2). Entropy captures the English word's average uncertainty, whereas surprisal captures how probable the particular form a learner must produce is. This parallels the translation-probability measure used in psycholinguistic work (Prior et al., 2013), estimated from parallel text as in Lim et al. (2023) rather than from elicited norms.

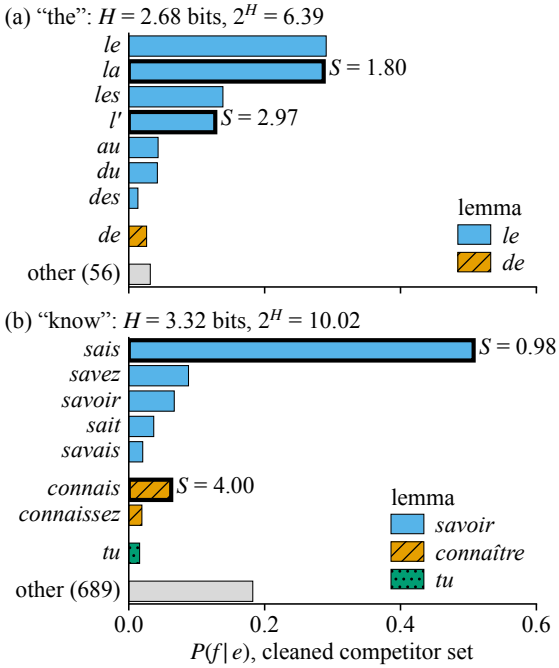
2.7 Comparison measures

We test whether surprisal is simply capturing other properties of the English word. Competitor count measures the number of distinct French forms in the cleaned competitor set (≥5 links). Contextual diversity (Adelman et al., 2006) is the number of distinct OpenSubtitles sentences containing the English word. English sense count is the number of WordNet synsets (Fellbaum, 1998). Because translation ambiguity can correlate with within-language polysemy (Degani et al., 2016), sense count tests whether surprisal is merely a proxy for English polysemy.

Finally, reverse entropy measures uncertainty over English words given the required French form, $H(E \mid f^*)$. If reverse entropy predicted error, the effect would be consistent with general translation uncertainty rather than the demands of producing French. All count measures are log-transformed.

2.8 Exposure variables

Four exposure measures are calculated from each learner's chronological record, using only events before the current token: prior productions of the required form, prior errors, days since last production, and a never-seen indicator. Prior productions are entered as $\log(1 + \text{count})$; days since last production is set to zero for unseen forms.



Prompt	Required form	surprisal (bits) Lexical	Inflectional	Total S
"the"	*la*	0.07	1.73	1.80
"the"	*l'*	0.07	2.90	2.97
"know"	*sais*	0.35	0.63	0.98
"know"	*connais*	3.16	0.85	4.00

Figure 2: Translation distributions and surprisal for two English prompt words. Each bar represents a French form coloured by lemma; bar length is $P(f \mid e)$ on the cleaned competitor set. Remaining forms are pooled as "other". *Tu* is alignment noise from *tu sais* ('you know'); it passes the part-of-speech filter as the past participle of *taire*. Table values are rounded independently, so Lexical + Inflectional may differ from Total S by 0.01.

2.9 Covariates

All models include log frequency of the French form (Lexique), log alignment count of the English word, token length, part of speech, sentence position, sentence length, days since course start, and the four exposure variables. Form frequency and English-word alignment count correlate at $r = .90$; neither coefficient is interpreted. Continuous predictors are z-scored once on the primary frame.

2.10 Model

Binary error is modelled with logistic mixed-effects regression. Crossed random intercepts for learner, word form, and exercise (Baayen et al., 2008) account for learner ability, word-form difficulty, and sentence-level difficulty, respectively. The word-form intercept is defined on the surface form rather than the lemma.

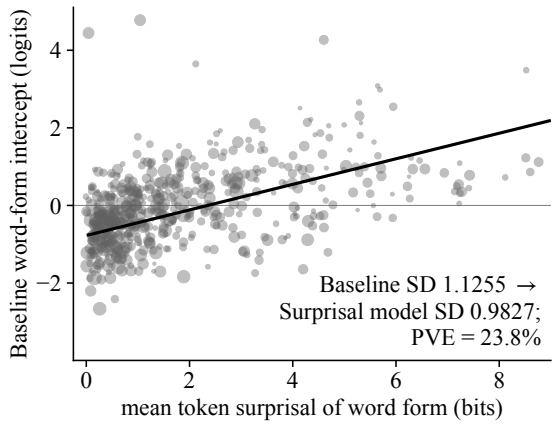


Figure 3: Representation of what PVE measures. Each point is one of the 593 word forms with at least 30 tokens (marker area increases with log token count); the line is a token-weighted linear fit. The y-axis represents difficulty in the Baseline model, after covariates, and the x-axis represents mean token surprisal. Harder words tend to be higher-surprisal words. Three word forms with mean surprisal above 9 bits lie beyond the axis and are included in the fit.

Model	Adds to Baseline
Baseline	Covariates + random intercepts for learner, word form, exercise
Count	Log competitor count
Entropy	Entropy of the translation distribution
Surprisal	Surprisal of the required form
Surprisal split	Lexical + inflectional surprisal
Entropy split	Lexical + inflectional entropy
Alternatives	Surprisal + competitor count + contextual diversity + English sense count
Practice	Surprisal + surprisal × prior productions + surprisal × days since last production
Course set	Entropy or surprisal (separate models) on the Course Set
Direction	Surprisal + reverse entropy

Table 1: Model specifications; each is the Baseline plus the listed terms.

The model containing only covariates and random intercepts is the Baseline. Named models add the predictors listed in Table 1. Because the word-form intercept captures residual between-word difficulty, adding a word-level predictor moves explained variation from the intercept into the predictor coefficient. We quantify this as the proportion of Baseline word-form intercept variance explained (PVE; Figure 3):

$$\mathrm{PVE} = \frac{\sigma^2_{w,0} - \sigma^2_{w,m}}{\sigma^2_{w,0}}, \tag{6}$$

where $\sigma^2_{w,0}$ and $\sigma^2_{w,m}$ are the word-form intercept variances in the Baseline and in model m.

We also report odds ratios (ORs). An OR of 1.62 means that a one-standard-deviation increase in the predictor multiplies the odds of error by 1.62, holding other model terms constant. Confidence intervals are 99% Wald intervals. Likelihood-ratio tests compare each model with the model it extends (Table 1).

The Baseline corresponds to a Rasch model, with learner intercepts representing ability and word-form intercepts representing difficulty; adding word-level predictors produces an explanatory item-response model (De Boeck and Wilson, 2004).

Models were fitted in R 4.6.1 with `lme4` 2.0.6 (Bates et al., 2015), using `glmer`, `nAGQ` = 0, and the nloptwrap BOBYQA optimiser. The Surprisal and Surprisal-split models were additionally refitted with the Laplace approximation (`nAGQ` = 1).

2.11 Cold-start evaluation

To test whether surprisal transfers to unseen vocabulary (a cold-start setting; McCarthy et al., 2021), word forms are split 80/20 five times using fixed seeds, with 337 held-out forms per split. The Baseline, Count, Entropy, and Surprisal models are trained on the 80%; held-out tokens are predicted using fixed effects and learner intercepts, without word-form intercepts.

We report token-level AUC, log loss, and the Spearman correlation between predicted and observed error rates for held-out word forms. We also estimate split-half reliability of observed error rates across learner halves; after Spearman-Brown correction, its square root provides an upper bound on the correlation attainable by a word-level predictor. Stratified splits by part of speech and frequency tercile are reported as a replication.

2.12 Robustness

Eight robustness checks test whether the surprisal effect depends on particular modelling choices. The lexical and inflectional components are permuted across word forms within frequency × length strata (200 replicates; word-level binomial model, deviance-drop statistic). The Surprisal-split model is refitted on random learner halves, within major parts of speech, and excluding determiners and pronouns. The Surprisal model is refitted excluding the 372 potentially mislinked tokens, with surprisal winsorised at the 99th percentile, using the Laplace approximation, and on the tile frame. Highly correlated predictor pairs are reported in

Appendix A.

3 Results

3.1 Descriptives and variance components

Baseline standard deviations are 0.73 for learner, 1.13 for word form, and 0.91 for exercise, indicating greater between-word than between-learner variation in error propensity.

3.2 Surprisal predicts error

Adding surprisal to the Baseline (Table 2) yields OR = 1.62 per standard deviation (99% CI [1.58, 1.66], $p < .001$; PVE = 23.8%). At the base error rate of 0.206, this corresponds to an increase in predicted error probability from 0.206 to 0.296.

3.3 Surprisal versus entropy and count

Competitor count (OR = 1.18, PVE = 3.4%) and entropy (OR = 1.17, PVE = 2.7%) both predict error ($p < .001$), but each explains substantially less between-word variance than surprisal. The distinction is particularly clear for inflection: inflectional entropy is weak or null (OR = 1.00; 1.10 with Lexique rather than UD determiner lemmas), whereas inflectional surprisal predicts error (OR = 1.31; Table 2). Thus, the probability of the required form is more informative than the word's average uncertainty.

3.4 Lexical and inflectional surprisal

Both components predict error: lexical surprisal OR = 1.47 [1.44, 1.50], and inflectional surprisal OR = 1.31 [1.26, 1.36], both $p < .001$. Lexical surprisal has the larger coefficient and together the two components account for 21.2% of between-word variance.

3.5 Alternative explanations

Adding competitor count, contextual diversity, and English sense count leaves the surprisal effect essentially unchanged (OR = 1.61 [1.57, 1.65]). Competitor count has no additional effect; contextual diversity is collinear with the alignment-count covariate ($r = 1.00$) and is not interpreted; English sense count retains a small effect (OR = 1.09 [1.05, 1.13]).

3.6 Direction

Reverse entropy does not increase error (OR = 0.91 [0.82, 1.01]), while the surprisal effect remains unchanged (OR = 1.62). This pattern is consistent with the effect being specific to the demands of producing French rather than reflecting general translation uncertainty.

3.7 Course Set

Course-set surprisal performs similarly to full-set surprisal (OR = 1.61 [1.57, 1.66], PVE = 26.5%; full set OR = 1.62, PVE = 23.8%). Course-set and full-set surprisal correlate at 0.98 across tokens, consistent with the median English word having 83% of its translation probability on forms in the course vocabulary.

3.8 Practice

Prior productions reduce error overall (Baseline OR = 0.65 [0.62, 0.67]) but do not reduce the surprisal gap. The surprisal $\times$ prior-productions interaction is 1.065 [1.046, 1.084], indicating that the surprisal effect increases slightly with practice. Days since last production does not moderate the effect (OR = 1.00 [0.99, 1.01]). Among tokens for which the learner had never previously failed the form, the interaction remains positive (OR = 1.046), suggesting that error-driven scheduling is not the sole explanation.

3.9 Cold start

For held-out word forms, adding surprisal increases AUC from 0.673 to 0.714, reduces log loss from 0.456 to 0.440, and increases the Spearman correlation between predicted and observed word difficulty from 0.18 to 0.37 (Table 6). Count and entropy each reach only 0.21. The result is stable across five splits (SD = 0.03) and stratified splits (0.35 versus 0.19 for the Baseline).

Observed held-out error rates have split-half reliability of 0.45, or 0.62 after Spearman-Brown correction, giving an approximate ceiling of $\sqrt{0.62} \approx 0.79$. On word forms observed in both learner halves, the Surprisal model reaches 0.41, or roughly half of this ceiling.

3.10 Sensitivity and tile exercises

The surprisal effect is stable across sensitivity analyses (Table 7). Its OR is 1.68 after excluding the 372 potentially mislinked tokens, 1.82 after winsorisation, and 1.71 with the Laplace approximation. Lexical and inflectional surprisal exceed the permutation null in 198 and 200 of 200 replicates, and remain positive under learner split-half, exclusion of determiners and pronouns (lexical

Model	Term	OR per SD	99% CI	PVE_{word}
Count	log competitor count	1.18	[1.13, 1.23]	3.4%
Entropy	entropy	1.17	[1.13, 1.22]	2.7%
Surprisal	surprisal	1.62	[1.58, 1.66]	23.8%
Surprisal split	lexical surprisal	1.47	[1.44, 1.50]	21.2%
Surprisal split	inflectional surprisal	1.31	[1.26, 1.36]	
Entropy split	lexical entropy	1.23	[1.19, 1.28]	−0.3%
Entropy split	inflectional entropy	1.00	[0.96, 1.04]	
Alternatives	surprisal	1.61	[1.57, 1.65]	24.3%
Alternatives	log competitor count	0.97	[0.92, 1.02]	
Alternatives	contextual diversity	0.86	[0.43, 1.72]	
Alternatives	English sense count	1.09	[1.05, 1.13]	
Practice	surprisal × log prior productions	1.065	[1.046, 1.084]	
Practice	surprisal × days since last	1.00	[0.99, 1.01]	
Course set	surprisal (Course Set)	1.61	[1.57, 1.66]	26.5%
Course set	entropy (Course Set)	1.14	[1.10, 1.18]	3.7%
Direction	reverse entropy	0.91	[0.82, 1.01]	

Table 2: Main models, primary frame (typed translation, 384,047 tokens). OR per SD of the standardised predictor with 99% Wald CI; PVE_{word} is the proportion of between-word variance explained relative to the Baseline. PVE is slightly negative when the word-form variance estimate does not shrink.

1.42, inflectional 1.31), and part-of-speech analyses. On tile exercises, estimates are similar to the primary frame: overall surprisal OR = 1.60 [1.54, 1.66], lexical surprisal = 1.45, and inflectional surprisal = 1.29.

4 Discussion

The results extend laboratory findings on translation probability to 384,000 learner productions. A corpus-derived measure of the probability of the required French form predicts learner error, despite being estimated from an unrelated parallel corpus. This supports the use of corpus-derived translation distributions where elicited translation norms are unavailable and allows the analysis to distinguish lexical from inflectional surprisal and to test production direction.

Several results narrow the interpretation of the effect. First, surprisal substantially outperforms competitor count and entropy. In particular, inflectional surprisal predicts error even though inflectional entropy barely does, indicating that the probability of the specific required form matters more than the English word's average uncertainty. Second, competitor count contributes no additional information once surprisal is included. Third, reverse entropy does not increase error, while production-direction surprisal does. Together, these findings are consistent with difficulty arising from competition among candidate translations weighted by their probability, although the correlational design cannot directly establish this mechanism.

Both lexical and inflectional surprisal contribute, with a larger lexical effect. The lexical result is consistent with accounts in which candidate translations compete according to their probability (Tokowicz and Kroll, 2007; Prior et al., 2013). The inflectional result indicates that probabilistic difficulty also extends to grammatical distinctions in the required form. The difference between the two results is also consistent with staged models of production where lexical selection is considered a competition between candidate lemmas, followed by the inflected form being built from the selected lemma and the sentence's grammatical features (Levelt et al., 1999): probability should be most relevant at the competitive first stage, and should matter less at the second stage, which relies on rules to determine forms. Some residual inflectional effect may reflect forms that beginners had memorised rather than derived.

Locating inflectional difficulty at the form level supports the use of a difficulty feature with separate lexical and inflectional components in CALL tasks, since a difficulty assigned per lemma cannot separate the probable and improbable realisations of the same lemma. Examples are learner modelling, where a learner producing one inflectional form does not imply the ability to produce others, and preparing assessments, where a lemma-level

difficulty measure will misjudge the difficulty of a specific required form.

Practice reduces errors overall but does not eliminate the difference between more and less probable forms. The surprisal effect increases slightly with prior productions and is unchanged by time since last exposure. Because Duolingo's scheduling is error-sensitive (Settles and Meeder, 2016), this interaction should not be interpreted causally. Separating learning from adaptive scheduling would require data from a system with non-adaptive exposure.

For ICALL, the cold-start result is particularly relevant. Surprisal is defined by a language pair rather than by the target word alone and can therefore provide a difficulty estimate before any learner has attempted the word. On held-out word forms, it substantially improves prediction over the Baseline. Unlike a purely predictive learner-history feature, it can be computed independently of prior learner performance and remains interpretable in terms of the translation distribution.

One application of the measure is to provide initial difficulty estimates for new items a course introduces. This set of initial difficulty scores then allows for ordering of items in the course such that probable forms are introduced earlier than improbable forms, while improbable forms are allotted more repetitions in the spacing schedule from the outset, rather than waiting for failures to accumulate. For lemmas already included in the course, the inflectional component can flag forms likely to be produced incorrectly before a learner has attempted them, so practice can focus on them early on.

A difficulty measure that separates lexical from inflectional difficulty is also useful in providing a system with an estimate of what type of error is more likely to arise from a token. This would allow a system to prepare feedback, whether it be a contrast between the meanings of competing words where lexical surprisal is high, or an agreement or conjugation rule where inflectional surprisal is high. Error-annotated learner data are necessary to check this against what types of errors learners actually make.

5 Limitations

The SLAM data consist largely of beginner vocabulary and cover only learners' first 30 days, so whether the effect generalises to higher proficiency is unknown. OpenSubtitles distributions also differ from language-learning materials; the Course Set analysis partly addresses this mismatch but does not remove it. Competitor sets remain sensitive to the mapping between UD and Lexique categories, although the required-form exemption protects the primary surprisal measure.

Exposure measures partly reflect Duolingo's error-sensitive scheduling, so practice interactions are descriptive rather than causal. The study also tests only English-French; although the measure can be computed for other language pairs with suitable parallel corpora, cross-linguistic generalisation remains untested.

6 Conclusion

We tested whether the probability of the French form an English speaker must produce predicts learner error, estimating this probability from parallel text without learner data. Across 384,000 Duolingo productions, surprisal predicts production error, with the effect concentrated on the probability of the required form rather than average translation uncertainty or the number of alternatives. Both lexical and inflectional surprisal contribute, and the effect is asymmetric: uncertainty in the reverse direction does not increase error.

The result identifies a language-pair-specific, interpretable difficulty feature that can be computed before a learner attempts a word. Surprisal also predicts unseen word forms substantially better than the Baseline, making it a potential feature for ICALL systems that estimate item difficulty for new vocabulary. The evidence is limited to beginner English-French learners and is correlational, so the proposed mechanism remains indirect. Future work should test additional language pairs, including the SLAM Spanish track, and use learner data annotated by error type to examine the mechanisms underlying surprisal-based difficulty more directly.

AI Disclosure

An AI assistant was used for parts of the analysis code and for language editing/polishing of our own text for clarity.

References

James S. Adelman, Gordon D.A. Brown, and José F. Quesada. 2006. Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.

David Alfter and Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88, New Orleans, Louisiana. Association for Computational Linguistics.

R.H. Baayen, D.J. Davidson, and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Aditi Chaudhary, Kayo Yin, Antonios Anastasopoulos, and Graham Neubig. 2021. When is Wall a Pared and when a Muro?: Extracting rules governing lexical selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6911–6929, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul De Boeck and Mark Wilson, editors. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York.

Tamar Degani, Anat Prior, Chelsea M. Eddington, Ana B. Arêas da Luz Fontes, and Natasha Tokowicz. 2016. Determinants of translation ambiguity: A within and cross-language comparison. *Linguistic Approaches to Bilingualism*, 6(3):290–307.

Tamar Degani and Natasha Tokowicz. 2010. Ambiguous words are harder to learn. *Bilingualism: Language and Cognition*, 13(3):299–314.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

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.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Johannes Graën and Gerold Schneider. 2017. Crossing the border twice: Reimporting prepositions to alleviate L1-specific transfer errors. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 18–26, Gothenburg, Sweden. LiU Electronic Press.

Elise Hopman, Bill Thompson, Joseph L. Austerweil, and Gary Lupyan. 2018. Predictors of L2 word learning accuracy: A big data investigation. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 511–516.

Jannika Laxén and Jean-Marc Lavaur. 2010. The role of semantics in translation recognition: effects of number of translations, dominance of translations and semantic relatedness of multiple translations. *Bilingualism: Language and Cognition*, 13(2):157–183.

Willem J. M. Levelt, Ardi Roelofs, and Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–38.

Zheng Wei Lim, Trevor Cohn, Charles Kemp, and Ekaterina Vylomova. 2023. Predicting human translation difficulty using automatic word alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11590–11601, Toronto, Canada. Association for Computational Linguistics.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1742–1748, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Arya D. McCarthy, Kevin P. Yancey, Geoffrey T. LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 883–899, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George A. Miller. 1955. Note on the bias of information estimates. In Henry Quastler, editor, *Information Theory in Psychology: Problems and Methods*, pages 95–100. Free Press, Glencoe, IL.

Boris New, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Anat Prior, Judith F. Kroll, and Brian MacWhinney. 2013. Translation ambiguity but not word class predicts translation performance. *Bilingualism: Language and Cognition*, 16(2):458–474.

Anat Prior, Brian MacWhinney, and Judith F. Kroll. 2007. Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods*, 39(4):1029–1038.

Anat Prior, Shuly Wintner, Brian MacWhinney, and Alon Lavie. 2011. Translation ambiguity in and out of context. *Applied Psycholinguistics*, 32(1):93–111.

Alexander Rich, Pamela Osborn Popp, David Halpern, Anselm Rothe, and Todd Gureckis. 2018. Modeling second-language learning from a psychological perspective. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 223–230, New Orleans, Louisiana. Association for Computational Linguistics.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858, Berlin, Germany. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

David Strohmaier and Paula Buttery. 2024. Semantic error prediction: Estimating word production complexity. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 209–225, Rennes, France. LiU Electronic Press.

Natasha Tokowicz and Judith F. Kroll. 2007. Number of meanings and concreteness: Consequences of ambiguity within and across languages. *Language and Cognitive Processes*, 22(5):727–779.

Natasha Tokowicz, Judith F. Kroll, Annette M. B. de Groot, and Janet G. van Hell. 2002. Number-of-translation norms for Dutch–English translation pairs: A new tool for examining language production. *Behavior Research Methods, Instruments, & Computers*, 34(3):435–451.

Brendan Tomoschuk and Jarrett Lovelett. 2018. A memory-sensitive classification model of errors in early second language learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 231–239, New Orleans, Louisiana. Association for Computational Linguistics.

Yun Wen and Walter J. B. van Heuven. 2017. Chinese translation norms for 1,429 English words. *Behavior Research Methods*, 49(3):1006–1019.

A Alignment, token linking and correlated predictors

Alignment. Word alignments were trained with fast_align (Dyer et al., 2013) in both directions on a uniform random subsample of 8,930,148 sentence pairs of OpenSubtitles v2018 en–fr (Lison and Tiedemann, 2016; Lison et al., 2018) (5 EM iterations, diagonal prior with optimised tension, Dirichlet prior on lexical distributions) and symmetrised with grow-diag-final-and, giving 62,057,793 links. $C(e, f)$ is the symmetrised link count; $P(f \mid e) = C(e, f)/C(e, \cdot)$ and $P(e \mid f) = C(e, f)/C(\cdot, f)$.

Token linking. Each SLAM French token f at normalised position pos_f is linked to the prompt word e at normalised position pos_e maximising

$$\begin{aligned}\mathrm{score}(e, f) = P(f \mid e)\, P(e \mid f) \\ \times \exp\big(-4\,|\mathrm{pos}_f - \mathrm{pos}_e|\big);\end{aligned} \tag{7}$$

tokens with all scores zero (5,136; 0.69%) are excluded.

Part-of-speech compatibility. Competitor forms are kept only if some Lexique category of the form is compatible with the prompt word's class, the majority UD part of speech of its linked SLAM tokens (Table 3).

SLAM UD POS class	Compatible Lexique cgram prefixes
NOUN	NOM
PROPN	NOM
VERB	VER, AUX
AUX	VER, AUX
ADJ	ADJ
ADV	ADV
DET	ART, ADJ:dem, ADJ:pos, ADJ:ind, ADJ:int, ADJ:num
PRON	PRO
ADP	PRE
CONJ	CON
SCONJ	CON
NUM	ADJ:num, NOM
INTJ	ONO
other	no POS filter

Table 3: Lexique categories (cgram prefixes) accepted for each SLAM UD part-of-speech class.

Count floor. The pre-specified rule (the smallest floor k at which $r(\log \mathrm{COUNT}_k, \log N) < 0.5$, otherwise $k = 5$) selected $k = 5$ by its otherwise-clause: the correlation rises with the floor (Table 4).

Correlated predictor pairs. Table 5 lists every pair of model terms with $|r| > 0.7$ on the primary frame.

floor k	words	$r(\log \mathrm{COUNT}_k, \log N)$	$r(\log \mathrm{COUNT}_k, \log \mathrm{CONTEXT_DIV})$	median COUNT_k
1	1558	0.521	0.537	90.000
3	1555	0.709	0.722	29.000
5	1552	0.752	0.764	19.000
10	1549	0.791	0.802	11.000
25	1538	0.825	0.833	6.000

Table 4: Count-floor diagnostic (English words with $N \geq 50$ alignment links).

Term	**Term**	r
form frequency (log)	English-word alignment count (log)	0.90
form frequency (log)	token length	−0.81
form frequency (log)	contextual diversity (log)	0.89
English-word alignment count (log)	token length	−0.72
English-word alignment count (log)	contextual diversity (log)	1.00
token length	contextual diversity (log)	−0.71
surprisal	surprisal (Course Set)	0.98
inflectional surprisal	inflectional entropy	0.74
entropy	lexical entropy	0.75
entropy	lexical entropy (Lexique lemmas)	0.86
entropy	entropy (Course Set)	0.86
lexical entropy	lexical entropy (Lexique lemmas)	0.76
lexical entropy (Lexique lemmas)	entropy (Course Set)	0.81
competitor count (log)	competitor count (Course Set, log)	0.92
contextual diversity (log)	competitor count (Course Set, log)	0.71

Table 5: Predictor pairs with $|r| > 0.7$ (primary frame, token level).

B Cold-start and robustness detail

Model	Split	AUC	log loss	Spearman
Baseline	random	0.673 ± 0.006	0.456 ± 0.019	0.177 ± 0.034
Count	random	0.676 ± 0.004	0.452 ± 0.019	0.210 ± 0.032
Entropy	random	0.685 ± 0.004	0.449 ± 0.019	0.208 ± 0.038
Surprisal	random	0.714 ± 0.006	0.440 ± 0.017	0.365 ± 0.029
Baseline	stratified	0.662 ± 0.028	0.505 ± 0.041	0.190 ± 0.055
Surprisal	stratified	0.702 ± 0.026	0.487 ± 0.043	0.353 ± 0.062
Split-half reliability	learner halves	–	–	0.45 (Spearman-Brown 0.62; ceiling 0.79)

Table 6: Cold-start evaluation on held-out French word forms (mean $\pm$ SD over five splits).

Check	Item	Value
permutation (200 replicates)	lexical surprisal	exceeds the null in 198 of 200 replicates
permutation (200 replicates)	inflectional surprisal	exceeds the null in 200 of 200 replicates
learner half A	lexical; inflectional surprisal	OR 1.488; 1.346 (n = 195,427)
learner half B	lexical; inflectional surprisal	OR 1.434; 1.325 (n = 188,620)
excluding determiners and pronouns	lexical; inflectional surprisal	OR 1.423; 1.307 (n = 243,998)
within part of speech (within-class SD)	NOUN: lexical; inflectional	OR 1.275; 1.372 (n = 90,506)
within part of speech (within-class SD)	VERB: lexical; inflectional	OR 1.553; 1.338 (n = 80,388)
within part of speech (within-class SD)	DET: lexical; inflectional	OR 1.479; 1.393 (n = 80,070)
within part of speech (within-class SD)	PRON: lexical; inflectional	OR 1.573; 1.037 (n = 59,979)
within part of speech (within-class SD)	ADJ: lexical; inflectional	OR 1.217; 1.280 (n = 32,279)
within part of speech (within-class SD)	ADV: lexical; inflectional	OR 1.322; 1.090 (n = 9,698)
winsorised at the 99th percentile	surprisal	OR 1.817 [1.762, 1.873]
winsorised at the 99th percentile	lexical; inflectional surprisal	OR 1.697 [1.653, 1.743]; 1.417 [1.353, 1.484]
excluding tokens above 12 bits	surprisal	OR 1.680 [1.632, 1.730]
excluding tokens above 12 bits	lexical; inflectional surprisal	OR 1.522 [1.487, 1.558]; 1.319 [1.267, 1.374]
Laplace approximation (`nAGQ` = 1)	surprisal	OR 1.71
tile frame (`reverse_tap`)	surprisal	OR 1.597 [1.537, 1.660]
tile frame (`reverse_tap`)	lexical; inflectional surprisal	OR 1.448 [1.405, 1.492]; 1.293 [1.207, 1.384]

Table 7: Robustness checks listed in Section 2.12: odds ratios per SD, with 99% CI where computed.