Semantic Error Prediction: Estimating Word Production Complexity*

David Strohmaier
ALTA Institute,
University of Cambridge
david.strohmaier@cl.cam.ac.uk

Paula Buttery
ALTA Institute,
University of Cambridge
paula.buttery@cl.cam.ac.uk

Abstract

Estimating word complexity is a well-established task in computer-assisted language learning. So far, however, complexity estimation has been largely limited to comprehension. This neglects words that are easy to comprehend, but hard to produce. We introduce semantic error prediction (SEP) as a novel task that assesses the *production* complexity of content words. Given the corrected version of a learner-produced text, a system has to predict which content words are replacements for word choice errors in the original text. We present and analyse one example of such a semantic error prediction dataset, which we generate from an error correction dataset. As neural baselines, we use BERT, RoBERTa, and LLAMA2 embeddings for SEP. We show that our models can already improve downstream applications, such as predicting essay vocabulary scores.

1 Introduction

Automatically estimating complexity of a word is a core task for computer-assisted language learning (CALL). This literature uses "complexity" to refer to the difficulty of processing a word (cf. North et al., 2023). But words can be difficult to process in multiple ways, leading to varieties of complexity. So far, the focus in NLP has been largely on complexity in comprehension. We fill a gap left by this focus and investigate the overlap of two varieties of complexity:

1. Lexical Semantic Complexity: The difficulty of a word due to its meaning.
2. Production Complexity: The difficulty of producing a word.

In the next section, we will discuss these types of complexity and their overlap in more detail, establishing their nature and relevance for CALL. To investigate the overlap, i.e. *lexical semantic production complexity*, we propose the task of semantic error prediction (SEP) and create an SEP dataset from an error correction dataset. Our method can be applied to other error correction datasets.

After describing the creation method for our dataset, we perform a Bayesian logistic regression analysis of candidate features for predicting semantic errors. We then provide SEP baseline results using BERT, RoBERTa, and LLAMA2 embeddings and compare them with the performance of the feature-based regressions. Finally, we use scores from the LLAMA2-based model for predicting the vocabulary scores of L2 learner essays with a Bayesian linear regression.

Our contributions are as follows:

1. We propose a new CALL task, semantic error prediction, which offers access to lexical semantic production complexity.
2. We present a method for creating SEP datasets from error correction datasets as well as a dataset created that way.
3. We provide results from transformer-based models for the SEP task.
4. We showcase the use of SEP models for the downstream application of predicting essay vocabulary scores.

The scripts required for creating the dataset are available online at https://github.com/dstrohmaier/semantic_error_prediction.

2 Types of Complexity

Word complexity, understood here as the difficulty of a word in processing, has many varieties. We develop the two overlapping types of complexity investigated in this paper and why they are important for CALL.

*This paper reports on research supported by Cambridge University Press & Assessment. We thank Chris Bryant for comments, advice, and provision of code, and all anonymous reviewers for their comments.

David Strohmaier and Paula Buttery. Semantic Error Prediction: Estimating Word Production Complexity.
Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024). Linköping Electronic Conference Proceedings 211: 209–225.

2.1 Lexical Semantic Complexity

Lexical semantic complexity is the complexity of a word due to its meaning. It can be distinguished from e.g. the syntactic complexity of a sentence or the orthographic complexity of a word form. Lexical semantic complexity has long been recognised as one of the main forms of lexical complexity, although its exact nature has been heavily debated (Cutler, 1983).

The notion of lexical semantic complexity can be compared with that of lexical sophistication, which is often understood as the use of low frequency vocabulary items (Laufer and Nation, 1995), although more detailed analyses have been put forward (Kim et al., 2018). We will find that, in some contexts, frequent words are difficult to produce, suggesting a difference between lexical sophistication and *contextual* lexical semantic complexity.

Similarly, lexical semantic complexity can be distinguished from other aspects of lexical complexity, such as morphological complexity. "abjure" is morphologically simpler but arguably semantically more complex than "theatergoer". In fact, we find in section 5 that character length appears negatively related to contextual semantic complexity.

Lexical semantic complexity poses deep challenges for CALL applications, as semantic nuances can be subtle and, thus, identifying semantically challenging words can be difficult. At the same time, semantic correctness is especially important for communication, more so than e.g. word order and subject-verb-agreement. We can understand other speakers even when their sentences violate multiple grammatical rules, but when they produce multiple semantically incorrect words, communication tends to break down.[1]

One reason that lexical semantic complexity has been difficult for CALL applications is that few ways exist to estimate it on the word-level. Compared to morphological and syntactic complexity, for which syllable count and depth of the syntactic graph serve as easily accessible features, features to predict the semantic complexity of a word token are harder to engineer. Many measures for assessing lexical complexity, such as the type-token ratio, operate on the document rather than the word-token-level (consider the features in table 1 of Bulté and Housen, 2012, p. 31).

There exist word-type-level features commonly associated with lexical complexity, such as:

- word frequency,
- age-of-acquisition (for first language speakers), and
- concreteness of the word.

As word-type-level features, they suffer from three shortcomings:

1. They ignore the contextual aspect of lexical complexity.
2. They typically fail to account for homonymy and polysemy, i.e. most data for them are only available on the word form level.
3. They cannot cover the entire vocabulary, as it rapidly evolves, e.g. how does the complexity of "rizz" compare to that of "mid"?

Hence, there is a need for another way to measure lexical semantic complexity in context, which we will meet.

2.2 Production Complexity

Production complexity, which we distinguish from comprehension complexity, is the difficulty of producing a linguistic unit either in speech or writing. For the purposes of the present investigation, production will be limited to writing.[2]

The distinction between comprehension and production complexity is related to the distinction between passive and active vocabulary, i.e. the recognition-recall difference, because production typically requires recall. Research into second language learning has investigated the difference, finding that even advanced learners show a large gap between passive and active vocabulary (Laufer, 1998; Fan, 2000).

Production complexity can diverge from comprehension complexity, because a semantic difference might be important for word choice without being important for word recognition. One example for this is the mass-count distinction. A language user might very well understand a sentence such as "He drank much milk." and yet erroneously produce sentences such as "He drank many milk.". That is, the mass-count distinction might play a bigger role in production than comprehension complexity.

[1] See Olsson (1972) and Khalil (1985) for support of the thesis that semantic errors impede communication more than grammatical errors. The research by Nushi et al. (2022) suggests that formal errors can reduce intelligibility more than lexical semantic errors, but, in their discussion, formal errors include e.g. the choice of the wrong suffix, which could arguably be treated as a semantic issue.

[2] For a survey of psycholinguistic research into task complexity and its interactions with other forms of complexity for L2 writing, see Johnson (2017).

For an example closely linked to word form, consider the case of "price"/"prize". In English, these two words differ in form and meaning. In German, however, the neargraph "Preis" is ambiguous between the two meanings. This might lead an L2 learner of English with German L1 to be able to comprehend the English words, while erroneously producing "prize" instead of "price".

Multiple data sources exist for assessing word complexity in general, with a tendency towards comprehension (Shardlow, 2013; Shardlow et al., 2020), while production is under-resourced.[3] One reason for this neglect is that much work on word complexity was intended to improve readability (see North et al. 2023, and, for an example, see the introduction of Gooding et al. 2021). Complex word identification was, thus, conceived of as a step in a pipeline for adapting text to a specific set of learners for *comprehension* (cf. North et al., 2023).

However, systems able to predict which words learners struggle to *produce* are also of use for adaptive teaching systems. Three such use cases are:

1. Content calibration: When learners are prompted to produce a particular word, the complexity of the word should be at the intended level for the task. For example, cloze tasks require learners to produce words that can fill a gap in a text. Knowing the production complexity of the target word would be of value for calibrating the item.
2. Assessment: Production complexity scores can serve to assess learner produced text, even though the relationship is not simple, as we will see in section 7.
3. Highlighting during learning: Words in a text read by a learner might be flagged to make the learner aware that they are harder to produce.

A further reason for the lack of resources on production complexity is that such datasets are harder to create. Eliciting complexity judgements from annotators reading a text is relatively simple. There does not appear to exist a simple equivalent for production, as it is challenging to ask annotators to rate the complexity of words while producing them at the same time.

To address this problem, we are using an error correction dataset based on learner-written texts for creating our SEP dataset.[4] Our method can be applied to any error correction dataset providing appropriate error annotations and corrections.

2.3 The Overlap: Lexical Semantic Production Complexity

We are interested in cases where a word is difficult to produce due to its semantics in a specific sentential context. This overlap gives rise to its own dynamics, because, in production, the conceptual information is typically activated prior to the word form information, rather than the reverse, as in the case of comprehension (see Jiang 2000 for an example of this). As a result of this reversal, we expect different complexity patterns in production than in comprehension.

Specifically, the patterns might show a different type of contextual effect: Language learners might inadvertently create contexts that require a certain word choice and as a result the learners might select the wrong word. Thus, a word that might be easy to comprehend and frequently selected in one context might be difficult to produce in another context, even though both contexts are created by the language user.

That lexical semantic production complexity is impacted by contextual effects is backed up by the empirical literature on English second language acquisition, which documents a sizeable number of semantic errors resulting from collocational phenomena (Al-Shormani and Al-Sohbani, 2012; Jeptarus and Ngene, 2016). Our approach and dataset provide a way for CALL applications to account for such phenomena specific to lexical semantic production complexity.

3 Related Work in NLP

Our work builds upon the NLP literature for both word complexity and error detection.

3.1 Word Complexity

The complex word identification (CWI) task, which has been investigated in multiple shared tasks (Paetzold and Specia, 2016; Yimam et al., 2017; Shardlow et al., 2021), aims to identify complex words in context. Recently, it has been extended under the name "lexical complexity prediction" (LCP) to a continuous task of predicting the complexity of a

[3] One resource specifically for production is the SweLLex word list (Volodina et al., 2016) for Swedish as an L2.

[4] Other options for estimating production complexity would include key-stroke or eye-tracking data. We thank an anonymous reviewer for these suggestions. These behavioural trace data, however, render it difficult to differentiate the semantic component of production complexity from other aspects.

word (Shardlow et al., 2021, 2020). For an in-depth review of the CWI/LCP literature, see North et al. (2023).

While feature-based machine learning system were state-of-the-art for many years (Gooding and Kochmar, 2018), by now end-to-end neural systems dominate the area (Shardlow et al., 2021). These models often use BERT-style transformers as their basis (Devlin et al., 2019; Liu et al., 2019). In the CWI literature, it has also been shown that backgrounds of language learners, e.g. their overall proficiency level, matter for whether a word is complex or not (Gooding et al., 2021).

As mentioned above, datasets in the CWI/LCP literature are generally more appropriate for capturing comprehension rather than production complexity. This tendency is due to the annotation process: annotators are presented with text for which they assign complexity labels. Effectively, the annotators engage in comprehension when deciding on a label.

Furthermore, complexity annotation is an artificial way of engaging with text, which raises the question of external validity. Even when the annotations are provided by L2 learners, theses learners are not trying to communicate with another human language user in a natural manner. By predicting errors in text production, our approach is closer to natural engagement with text and, therefore, addresses this issue.

There also exists a literature on predicting the CEFR levels of words (Alfter and Volodina, 2018; Pintard and François, 2020), which is less comprehension focused. This literature tends to consider words or word senses in isolation, rather than in the context of use (but see Aleksandrova and Pouliot, 2023).

3.2 Error Detection

Semantic errors are covered by the error detection literature, but much of this literature is focused on morpho-syntactic errors. Similar to complexity, error detection and closely related problems have been the subject of multiple shared tasks (Ng et al., 2014; Bryant et al., 2019; Volodina et al., 2023). Similar to CWI/LCP, this field is dominated by transformer-based models, often combined to increase performance (Qorib et al., 2022; Qorib and Ng, 2023). For a recent survey of error correction, see Bryant et al. (2023).

While closely related to SEP, error detection and correction systems are not designed for the purpose of assessing the lexical complexity of content words, but rather their correctness. Correctness, however, can be due to the learner avoiding more difficult terms and resorting to simpler expressions. By contrast, our approach is able to distinguish two correct words with regard to which one was more complex to produce.

4 Dataset

We present a SEP dataset that can be constructed from existing resources.[5] Our dataset uses error correction as the starting point for determining production complexity. Using learner texts as the source of the dataset ensures high external validity: The learners are engaged in a naturalistic task and patterns of their output are used to assess the lexical semantic complexity.

In constructing our dataset, we only predict the corrections of word choice errors. That is, we focus on the word tokens learners *should* have produced, but failed to do so. This production failure is taken as a direct indicator of production complexity.

Our approach only considers the corrected token, not the erroneously produced words. That is, when an annotator tags replaces "work" by "job" in a sentence, this is taken as evidence that the "job"-token in this sentence is complex, without any further inference regarding "work".

The reason for this choice is that we are interested in the complexity of word tokens in a specific context. It is unclear what we learn from an erroneously produced token. When a token is produce, it was evidently feasible to wrongly produce "work", even it it was semantically impossible to produce this word token correctly. Due to these conceptual problems, our dataset construction will focus on the context-appropriate words that learners fail to produce.

Our dataset concerns both the breadth and depth of lexical knowledge (Bulté and Housen, 2012): Errors occur both when learners lack items in the vocabulary, an issue of breadth, and when learners lack the lexical knowledge to correctly integrate words into sentences, an issue of depth. Thus, our research cuts across the theoretical constructs of lexical complexity presented by Bulté and Housen (2012, figure 3).

[5]The scripts required for doing so are made available at: https://github.com/dstrohmaier/semantic_error_prediction.

4.1 Dataset Construction

Our starting point is the dataset published as part of the 2019 BEA shared task on grammatical error correction (Bryant et al., 2019). This dataset provides error annotations for sequences, taken primarily from texts written by second language learners of English, although the evaluation data also includes some native speakers. The annotations follow the scheme of the ERRANT tool (Bryant et al., 2017). In addition, the dataset provides CEFR levels for the texts (CoE, 2020).

Since we are interested in semantic word choice and such word choice can be evaluated only in a semantic context, we use whole paragraphs extracted from the dataset as input.[6] We then invert the dataset so as to move from error detection to error prediction.[7]

Error Code	Meaning	Example
R:VERB	Verb replacement	*order* → *book*
R:NOUN	Noun r.	*base* → *foundation*
R:ADJ	Adjective r.	*low* → *poor*
R:ADV	Adverb r.	*graciously* → *gracefully*

Table 1: Selected error types (cf. Bryant et al., 2017).

In the next step, we select the relevant error types: word replacement errors in which content words, i.e. nouns, verbs, adjectives, and adverbs have been replaced by the annotators.[8] Orthographic, morphological, tense, subject-verb agreement and similar are thus excluded from the prediction task to focus on semantic complexity. They are also corrected, however.

In addition, we render the labels binary: Each token in the dataset is annotated for whether it has been corrected using one of the selected error tags.[9]

[6]Extremely short paragraphs, for example best wishes at the end of a letter, are merged into larger paragraphs when possible. In a small number of cases, the sub-tokenized paragraphs are longer than the maximal sequence length (512). 7 texts are affected, only one of which belongs to the split used for evaluation.

[7]We thank Chris Bryant, one of the original organizers of the 2019 BEA shared task, for providing code.

[8]When the replacement crosses part of speech, e.g. a verb is replaced by a noun or an adverb by an adjective, Errant typically treats this as an R:OTHER error, which is not used by us. We assume that when such errors occur, usually more has gone wrong than just the choice of a wrong word due to its semantics.

[9]We only label word tokens with the spaCy POS-tags: VERB, NOUN, ADJ, ADV. As a result, we exclude a small number of positive labels for other POS. The largest block of these positive labels are auxiliary verbs. 344 out of 49118 are labelled positively, most of which are in their turn forms of "be", "have", and "do". We use the spaCy `en-core-web-sm` model for POS-tagging (Honnibal et al., 2020).

For evaluation of these binary tags, we use the F_1-score and the area under the curve (AUC) of the Receiver Operating Characteristic Curve (ROC).

The original dataset provides a public train- and a dev-split.[10] We use the dev-split as an eval-split and split the train-split into a new train- and new dev-split. We apply our method to these public splits, with the new dev-split being primarily used for development purposes prior to evaluation (e.g. checking code correctness).

	# sequences	# tokens	# content t.	% errors
train	10523	577892	239156	2.45
dev	1170	63420	26409	2.43
eval	1419	88580	36923	2.04

Table 2: Descriptive dataset metrics. "content t." stands for tokens with content word POS tags. Percentages indicate the percentage of content word tokens corresponding to replacement errors.

4.2 Descriptive Metrics

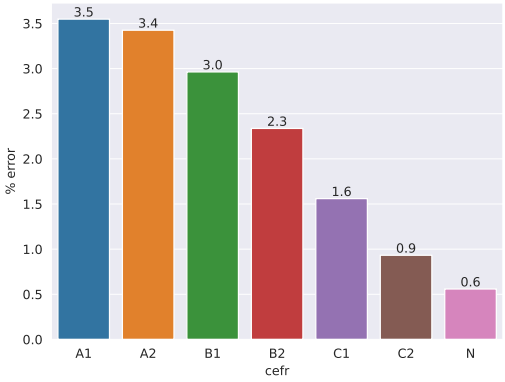


Figure 1: Replacement error percentages for content word tokens across CEFR levels (N=native speakers).

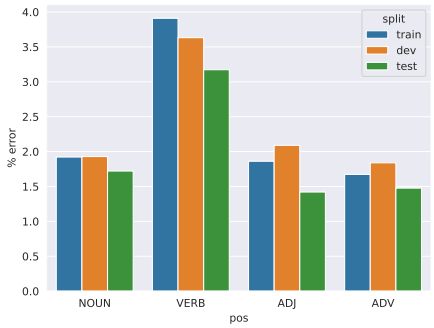
The training-split contains more than half a million tokens, slightly less than half of which are content word tokens (see table 2). The dev- and eval-split are $> 10\%$ of that size.

Across splits, around 2.4% of content word tokens correspond to semantic errors,[11]. These overall numbers, however, mask considerable differences in the error percentages across CEFR levels: The lower the CEFR level, the more content word

[10]The test split is not public and therefore not used by us.

[11]The number for the eval split in table 2 is lower, because the dev split also includes native speakers.

	POS	%
Train	Noun	1.9%
	Verb	3.9%
	Adj.	1.9%
	Adv.	1.7%
Dev	Noun	1.9%
	Verb	3.6%
	Adj.	2.1%
	Adv.	1.8%
Eval	Noun	1.7%
	Verb	3.2%
	Adj.	1.4%
	Adv.	1.5%



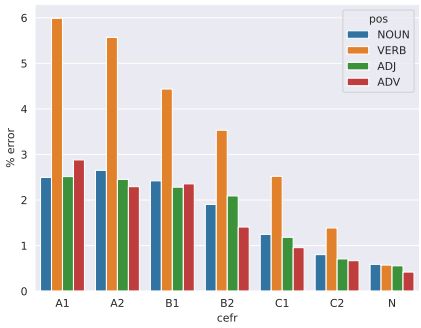


Figure 2: Percentages of errors for different POS across splits and CEFR levels (N=Native).

replacement errors are committed by a learner (see figure 1).

Across POS-tags, verbs are particularly likely to have been corrected, around 3.5% of the time (see figure 2). Verbs are a greater source of errors for second language learners of English, even at the C2 level, but the effect disappears for native learners of English, which are included in the eval split of the dataset. We speculate that this might be due to the higher context dependence of verbs, at least when compared to nouns (Gentner and France, 1988; Kersten and Earles, 2004; Earles and Kersten, 2017).[12] The context-dependence might take a language specific form, leading to L1-interference for L2 learners.

4.3 Qualitative Inspection

In this section, we consider hand-picked examples of replacement errors from our new train split.

The qualitative inspection suggests that learners often replace words with neargraphs, e.g. using "aspects" instead of "respects" or "affection" instead of "infections".[13] That being said, mistaken words and their corrections are also semantically related, with learners using "blame" instead of "guilt" and "contaminated" instead of "polluted".

Some mistaken tokens exhibit a lack of specificity. For example, in the corrected sentence "And some buses drive at night to transport [take → transport] passengers." "transport" replaces "take". While the meaning of the sentence can be understood without this correction, *transport* is more specific than *take*. It would be too simplistic, however, to think that learners always use less specific words.

Adjectives provide evidence that the words learners fail to produce are not necessarily highly specific or generally lacking from their vocabulary: "good" is one of the adjectives most often inserted by annotators. It typically replaces more specific adjectives such as "suitable" and "healthy" that fail to be contextually appropriate. This observation underlines the difference between lexical semantic complexity in production and comprehension: a learner producing "suitable" instead of "good" is likely able to comprehend the word "good".

A lack of idiomaticity can also lead to corrections. For example, annotators changed "big enterprises" to "big businesses". Similarly, annotators replace "main friend" with "best friend". In these cases, a reader will be able to comprehend the sentences with either word choice, but the corrected formulation is more idiomatic.

Errors due to a lack of idiomaticity are one reason why semantic error prediction is a highly challenging. Consider the following sentence:

"I began doing this sport three years ago when I lost my job [work → job]."

In this sentence, the annotators replaced "work" with "job", but this is a very nuanced correction, that arguably involves collocational preference as well as semantic detail.

5 Bayesian Regression Analysis

To analyse the dataset, we perform a Bayesian logistic regression using Bambi (Capretto et al., 2022), a package for Bayesian regression models based on PyMC (Oriol et al., 2023). We fit the regression on

[12] The claim that verbs are more context-dependent is related to the idea that a verb predicates something of something else, thus being constrained both by what it predicates and the subject of its predication. The idea that verbs play this connecting role might be tracked back at least to Aristotle, who in *De Interpretatione* (3.16b6–7) asserts that "it [a verb] is a sign of things said of something else" (Aristotele, 1994, p. 44).

[13] Errant provides a separate tag for orthographic error (R:ORTH), which we do not use.

the combined training and evaluation splits of the data using only tokens that were tagged as content words using spaCy (Honnibal et al., 2020). Since we are also primarily interested in interpreting the features, we drop rows that lack a feature required for any of the regression models.

We estimate two models. The first is a base model which has as input features (see next section for details):

- length of the word in characters,
- word frequency,
- age of acquisition, and
- whether the token is a verb.

The second models adds an interaction effect between being a verb and the frequency. The equations are described in appendix figure 6.

5.1 Observed Variables for Regression

Except for being a verb, all the explanatory variables were selected based on their general usage in the complex word identification literature (e.g. Gooding and Kochmar, 2018). However, in SEP the features are for the *corrected* learner text.

Length in characters. Provides the length of the token as counted in characters.

Frequency. We use the wordfreq package for Python,[14] specifically the Zipf frequency estimate. The package uses 0 as the default value of words not found in the word list.[15]

Age of acquisition (AoA). While the age of acquisition is a metric for L1 acquisition, it can also be applied to L2 acquisition under the simplifying assumption that both acquisition processes proceed from simpler to more complex words. While this assumption is probably not correct in all cases due to vocabulary transfer from L1 to L2, it offers a sufficiently close approximation of learning order (as our results show; see also the correlation of learning order documented by Flor et al. 2024). The AoA values are taken from the dataset presented in Kuperman et al. (2012).[16] The coverage by this dataset is incomplete and tokens for which no AoA is available are dropped from the dataset. Other tokens from the same sentence are still used for training and evaluation. We scale the age of acquisition to a mean of 0 and variance of 1 to make it comparable to the CEFR-j.

CEFR-j. The CEFR-j project provides the CEFR level of word types based on the word lists provided by Open Language Profiles and Octanove.[17] We convert and scale the CEFR-j data to make it comparable with the AoA.

Is verb. The spaCy tags were used for this feature. It was motivated by our previous analysis, suggesting that verbs are much more likely to be semantic errors (see section 4.2).

CEFR. The underlying dataset provides the CEFR level for the submissions. We treat this as a categorical variable.[18]

5.2 Results and Interpretation

A Bayesian logistic regression produces a probability distribution over the parameters of interest. For the estimated parameters, we report the highest density interval (HDI), i.e. the interval of minimum width containing the parameter with a certain probability. As is the standard for Bambi, we consider the 94% HDI credible interval (i.e. the interval spanning from 3% to 97%). HDIs are often treated analogously to frequentist confidence intervals, but have the straightforward interpretation that, given observed data,[19] the effect has a 94% probability of falling within the interval.

The results for the base model can be seen in table 3 and figure 3. In line with the expectations from the CWI literature and the previous analysis, we find that;

- more frequent content words are less likely to be semantic errors (HDI: $[-0.21, -0.11]$),
- content words with a higher CEFR level are more likely to be semantic errors (HDI: $[0.09, 0.16]$),

[14] https://github.com/rspeer/wordfreq. The packages uses the ExquisiteCorpus (https://github.com/LuminosoInsight/exquisite-corpus).

[15] 0 does not correspond to zero occurrences due to the zipfian transformation.

[16] Downloaded from https://osf.io/kz2px/.

[17] The lists were downloaded from https://github.com/openlanguageprofiles/olp-en-cefrj/. The CEFR-J Wordlist Version 1.5 was compiled by Yukio Tono, Tokyo University of Foreign Studies (Negishi et al., 2013). We use the CEFR-j list over others, because it is on the level of word form + POS rather than word sense. For example, the online EVP lists CEFR levels A1 and C2 among others for different senses of the noun "head", while CEFR-j only provides A1 for the noun. Furthermore, CEFR-j provides a permissive license and easy access.

[18] In contrast to CEFR-j, we do not convert and scale the CEFR-level to be able to compare it to the AoA. The reason for this difference is that we do not intend a comparison with AoA, because it is a student-level rather than token-level variable.

[19] More rigorously, given the model specification, the prior, and the observed data.

	mean	sd	$hdi_{3\%}$	$hdi_{97\%}$
is verb	0.80	0.03	0.75	0.85
cefr[C2]	-1.49	0.08	-1.6	-1.35
cefr[C1]	-0.91	0.06	-1.02	-0.8
cefr[B2]	-0.5	0.05	-0.59	-0.4
cefr[B1]	-0.25	0.05	-0.33	-0.15
cefr[A2]	-0.07	0.05	-0.15	0.03
scale(cefr-j)	0.12	0.02	0.09	0.16
scale(aoa)	0.09	0.02	0.05	0.13
frequency	-0.16	0.03	-0.21	-0.11
character length	-0.06	0.01	-0.07	-0.04
intercept	-2.39	0.16	-2.69	-2.08

Table 3: Estimated parameters of base model. Mean, standard deviation, and HDI boundaries of the estimated posterior are provided.

	mean	sd	$hdi_{3\%}$	$hdi_{97\%}$
frequency:is verb	0.22	0.04	0.15	0.28
is verb	-0.3	0.19	-0.64	0.06
cefr[C2]	-1.49	0.08	-1.63	-1.35
cefr[C1]	-0.91	0.06	-1.02	-0.81
cefr[B2]	-0.50	0.05	-0.60	-0.40
cefr[B1]	-0.25	0.05	-0.33	-0.15
cefr[A2]	-0.07	0.05	-0.16	0.02
scale(cefr-j)	0.12	0.02	0.09	0.16
scale(aoa)	0.10	0.02	0.06	0.13
frequency	-0.28	0.03	-0.34	-0.22
character length	-0.06	0.01	-0.07	-0.04
Intercept	-1.80	0.19	-2.17	-1.46

Table 4: HDI (3–97% interval) of model with interaction between being a verb and frequency (B+I). Table includes mean, standard deviation, and HDI boundaries of the posterior.

- content words with a higher AoA are more likely to be semantic errors (HDI: $[0.02, 0.05]$), and
- students with higher CEFR level are generally less likely to commit errors,
- verbs are more likely to be semantic errors (HDI: $[0.75, 0.85]$) compared to other content words.

Contrary to what one might expect from the CWI literature, longer words appear less likely to be replacements for semantic errors (HDI: $[-0.07, -0.04]$). That is, a word in the corrected sentence being longer is not an indicator of it corresponding to a semantic error. This could be a result of human annotators avoiding complex corrections for pedagogical reasons, or an effect of learners rarely intending to write long words. The finding is in line with "good" being one of the most frequent corrections for adjectives, replacing words like "suitable" and "healthy".

The second model, which introduces an interaction between frequency and being a verb, complicates the picture considerably (see table 4 and figure 4). Being a verb stops being a strongly positive predictor for semantic errors (HDI: $[-0.64, 0.06]$), while the interaction between frequency and being a verb is positive (HDI: $[0.15, 0.18]$). This additional analysis suggests that the effect of verbs being more likely to be semantic errors is due to *frequent* verbs. This is line with our speculation that verbs enjoy a special status due to their contextual dependence: learners struggle with verbs because they are heavily constrained by context, not because they are rare.

One practical implication of this result is that a focus on the correct use of frequent verbs could be beneficial to support learners in production.

6 Deep Learning Models

We put forward baseline deep learning models trained for semantic error prediction using our dataset. The models are probes trained on embeddings from pre-trained transformer models (Vaswani et al., 2017).

6.1 Architecture

We use the English BERT and RoBERTa models as the basis of our architecture (Devlin et al., 2019; Liu et al., 2019).[20] In addition to these well-researched models, we also explore the more recent and larger LLAMA2-7B model (Touvron et al., 2023).

We use these models without fine-tuning to create embeddings of the word tokens in question. Due to subtokenisation, the base model might produce more than one embedding per word token, which we address with mean pooling of the subtoken embeddings. Research has suggested that the last layers of BERT-like transformer models are not best suited for lexical semantic tasks (Vulić et al., 2020). Therefore, we create our embeddings by mean pooling over layers 1–10 (inclusive), i.e. excluding the last two layers, for the BERT and RoBERTa models. For LLAMA2, the role of the different layers is not as well-established and we resorted to averaging the output of layers 1–30, i.e. ignoring the last two layers again.

Probes are fine-tuned on the word embeddings, which requires less computational resources than training the whole transformer. The probes consist of a hidden layer (size 100), an output layer, and a SoftMax pooling layer, described by the following

[20] We use the base-size models. All models are loaded using the huggingface `transformers` library (Wolf et al., 2020).

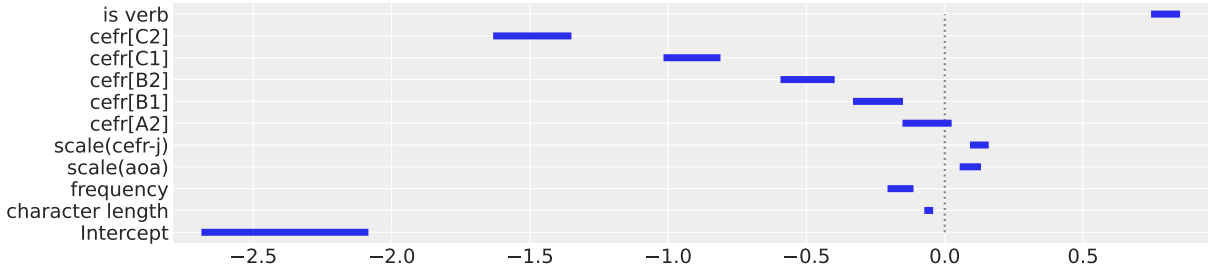


Figure 3: HDI credible intervals (3–97%) for coefficients of the base model (B).

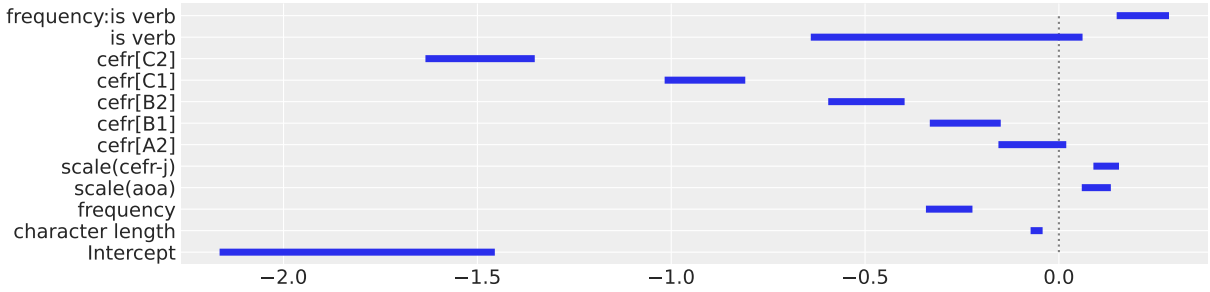


Figure 4: HDI (3–97% interval) for coefficients of the interaction model (B+I).

equations:

$$e = \text{embedding}$$
$$h = \text{ReLU}(W_{\text{hidden}} e + b_{\text{hidden}})$$
$$scores = \text{LogSoftMax}(W_{\text{output}} h + b_{\text{output}})$$

6.2 Training

The probes are trained by first performing hyperparameter-search using 5-fold cross-validation on the combined data from the train- and dev-split. The hyperparameter search randomly draws 20 hyperparameter settings from the space (see table 8 for details). The probe is then trained on the combined train- and dev-split using the hyperparameters reaching the highest F_1-score in cross-validation.[21] For training, we use the AdamW algorithm (Loshchilov and Hutter, 2018). The evaluation occurs on the eval-split.

The extreme label imbalance of the dataset can lead the probes to exhibit a bias towards assigning negative labels. To address this, we over-sample the positive labels during training, so that there is an equal number of positive and negative labels.

6.3 Results and Interpretation

Due to the imbalance of the labels, accuracy is not a meaningful metric for our dataset. Instead, we use the F_1-score and the area under the curve of the ROC (AUC). The AUC can be interpreted as the probability that a randomly chosen positive instance, i.e. a content word token that is a replacement, will have a higher score than a randomly sampled negative instance.

Table 5 provides the overall results, as well as the results for each CEFR level (including the "N" level for native speakers). We compare the deep learning models against a baseline that labels all tokens as corresponding to semantic errors ("all True"),[22] and the regression models discussed in section 5.[23]

With a threshold of 0.5, the logistic regression models fail to achieve an F_1-score of above 0%. The AUC score is more promising, consistently outperforming the 50%-threshold of the all True baseline. The transformer-embeddings based models outperform the regression baselines: with only one exception, there is at least one transformer model that outperforms the best regression model for every CEFR level. The exception is the AUC for the A1 level (B: 67.6%). In this case the information of the student CEFR level might be of sufficient importance to outweigh the performance

[21] The hyperparameter search space and the best hyperparameters for each probe are available in the online materials at https://github.com/dstrohmaier/semantic_error_prediction/tree/main/probe_kwargs.

[22] Labelling all tokens negatively would lead to an F_1 of 0.

[23] For the native test data, the CEFR label of the student is given as C2, since this is the closest available class. Otherwise the comparison to the regressions models favours the later, because they are not evaluated on missing data, e.g. when the age of acquisition of a word is not accessible.

	Overall		A1		A2		B1		B2		C1		C2		N	
	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC
all True	4.0	50.0	6.4	50.0	6.8	50.0	6.0	50.0	4.7	50.0	3.6	50.0	2.3	50.0	1.1	50
B	0.0	68.7	0.0	**67.6**	0.0	60.1	0.0	59.8	0.0	62.6	0.0	55.6	0.0	69.4	0.0	54.7
B+I	0.0	69.0	0.0	66.7	0.0	60.4	0.0	60.3	0.0	62.4	0.0	58.4	0.0	68.7	0.0	54.0
BERT	9.6	67.8	10.4	64.3	13.1	**69.1**	15.2	66.8	9.4	66.5	7.3	**68.2**	8.5	73.2	2.7	**59**
RoBERTa	10.8	69.2	**12.4**	64.7	13.8	67.2	13.0	70.1	**14.1**	**72.8**	**10.5**	66.3	**10.8**	**78.1**	1.8	**59**
LLAMA2	**11.0**	**69.8**	11.9	64.6	**14.9**	68.1	**15.7**	**70.8**	12.4	68.8	6.5	67.4	3.0	72	**2.8**	58.2

Table 5: Scores in percentages. The baseline scores result from assigning True to all tokens or all content word tokens.

advantage of the transformer embeddings.

Looking across CEFR levels, no simple trend in performance holds. Both A1 (highest transformer AUC: 64.7%) and C1 (highest AUC: 68.2%) appear particularly challenging. One generalisation that can be made is that the numbers on the native data are the worst ($F_{1:}$: 2.8%, AUC: 59%). We assume that this is due to the absence of native essays in the training data. In effect, this result strongly suggests that the error patterns for native and L2 speakers differ considerably. After all, the C2 level, which is supposedly the closest to the native skill, has the highest performance! That being said, the native data are from a different source, the LOCNESS corpus (Granger, 1998), which might also explain the low performance.

That LLAMA2 has the highest overall F_1 (11%) and AUC (69.8%) suggests that the size of the language model is a factor. Generally, however, the differences are small and the highest AUC value is achieved by RoBERTa for the C2 level (78.1%).

In light of the dataset difficulty, it is not surprising that the F_1-scores are low. The higher AUC are somewhat encouraging, especially for certain CEFR levels (e.g. C2 for RoBERTa reaching 78.1%). To support educational technologies, it will be important to better differentiate between complex and other words, i.e. to increase the AUC. That being said, the current scores can already be used as an input feature for downstream tasks, as we show in the next section.

7 Downstream Application

We show that the scores of one of our models support essay score prediction as a downstream task.

7.1 Setup

We use the ELLIPSE dataset (Crossley et al., 2023) for evaluation, which provides vocabulary scores for more then 6000 essays by L2 learners of English. We use the probability scores produced by our LLAMA2-embeddings based model as it is the overall best performing model (see table 5) to predict these vocabulary scores using a Bayesian linear regression.

The vocabulary scores are on the essay-level, while our lexical complexity scores are on the token level, requiring us to perform pooling. We consider two forms of pooling: mean and max pooling.

In addition, we compare the regression using our model-derived lexical complexity scores with a simpler approach: For the simple approach, we use the proportion of times a word has been put forward as a correction. We use again mean and max pooling.

We also include other variables that can be used to assess vocabulary in our regression:

Min. Frequency. We use the same source of word frequencies as discussed in section 5.1. We apply min-pooling to the token frequencies, removing frequencies of 0.0 (default value).[24]

CEFR-j. We use the CEFR-j word list discussed in section 5.1, applying min-max-normalisation, so that each CEFR level to corresponds to a 0.2 step, providing a range from 0–1 for comparison with our probe scores, which also range from 0 to 1. The CEFR-j scores for tokens are mean-pooled.

Type-Token Ratio. Following the literature on complexity (Bulté and Housen, 2012), we use the type-token ratio as a feature. The data is provided by the dataset, but we use the ratio rather than the percentage for comparability.

Measure of Textual Lexical Diversity (MTLD). The ELLIPSE dataset also provides MTLD data, a metric from lexical diversity derived from the type-token ratio (McCarthy, 2005), but accounting for text length. We rescale this data to a mean of 0 and standard deviation of 1.

[24]We also explored mean-pooling but found its coefficient to be indistinguishable from 0.

Grade Level. The ELLIPSE dataset includes students from grade 8 to 12. We use this information and min-max normalise the grade level to make it comparable with our probe scores.

We compare five regressions models:

1. **base**: Base model without any of our lexical semantic production complexity scores.
2. **max**: Model using the max-pooling of our lexical complexity scores in addition to base variables.
3. **mean**: Model using only the mean-pooling of our lexical complexity scores in addition to base variables.
4. **max+mean**: Model using both complexity scores.[25]
5. **proportion**: Model using the mean and max pooling error correction proportions instead, as described above.

7.2 Results and Discussion

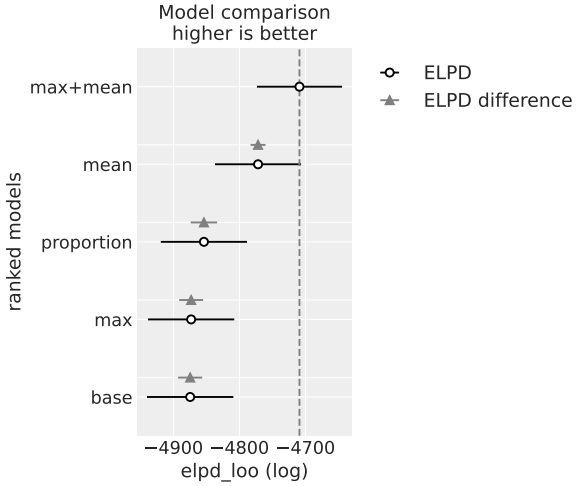


Figure 5: elpd_{loo} scores for Bayesian linear regression models predicting vocabulary scores.

To compare our five models, we use the expected log pointwise predictive density, which is estimated using leave-one out cross-validation (elpd_{loo}). The elpd_{loo} is a standard metric for comparing Bayesian models (see figure 5 and table 7) and can be written as (see Vehtari et al., 2017):

$$\text{elpd}_{\text{loo}} = \sum_{i}^{n} \int p(y_i|\theta)p(\theta|y_{-i})d\theta$$

where y_{-i} are all datapoints except the i-th.

[25]It might appear more appropriate to use the median rather than the mean, as the latter also incorporates the max value. We found, however, that this made a negligible difference.

	mean	sd	$\text{hdi}_{3\%}$	$\text{hdi}_{97\%}$
intercept	3.10	0.07	2.98	3.24
vocabulary σ	0.50	0.00	0.49	0.51
grade level	-0.15	0.02	-0.18	-0.12
max scores	0.71	0.06	0.59	0.83
mean CEFR-j	3.76	0.22	3.37	4.19
mean scores	-4.46	0.24	-4.92	-4.01
min frequencies	0.05	0.01	0.04	0.07
scale(MTLD)	0.20	0.01	0.18	0.21
type-token ratio	-1.23	0.09	-1.40	-1.06

Table 6: Results of Max+Mean model for predicting the vocabulary scores of ELLIPSE essays.

We also provide the R^2 metric in table 7 in the appendix, because it is more established within in NLP literature, although it neglects the probabilistic information provided by the Bayesian approach. It shows the same picture as the elpd_{loo} for the five models.

The comparison suggests that adding both the mean- and the max-pooled scores contribute to the fit of the model. The max-pooling, however, contributes only substantially when combined with the mean-pooling. The max+mean model also outperforms the proportion model, showing that the neural models are helpful.

We provide the HDI for our best fitting model in table 6 and figure 7. Among the features, the mean pooled score of our model has the largest absolute coefficient.[26] The coefficient is, however, negative (HDI: $[-4.92, -4.01]$), which might appear surprising at first glance. After all, a higher score should indicate more complex words, which in turn one might expect to indicate a higher proficiency. We believe that this puzzle can be explained by also taking into account the effect of the max-pooled scores.

The effect of the max-pooled scores is smaller, but with high probability positive (HDI: $[0.59, 0.83]$), thus pointing in the expected direction. We interpret this suggestion as follows: the skilled learner produces few contexts that might easily lead to confusion, thus rending the average word token easier to choose, but their most complex word is more challenging than that of a learner at a lower level.

The surprising negative coefficient is not just present for our scores, but also for type-token ratio[27] (HDI: $[-1.40, -1.06]$) and grade levels of

[26]No direct comparison to frequencies or the scaled MTLD is possible due to the different scale.

[27]The MTLD, however, has the expected relationship, sug-

students (HDI: $[-0.18, -0.12]$). In the case of the minimum frequencies, we find a somewhat surprising positive coefficient (HDI: $[0.04, 0.07]$), suggesting that a higher vocabulary score is associated with avoiding very rare words. De Wilde (2023) has previously found for L2 writing that "more proficient learners use more frequent words" (p. 11), but also notes that the literature is divided on this. These inverted results suggest a non-linear relationship between L2 learner writing and features which the literature associates with lexical sophistication.

	elpd_{loo}	$\text{elpd}_{\text{diff}}$	se	dse	R^2
max+mean	-4709.1	0.0	64.4	0.0	0.26
mean	-4771.8	62.7	65.2	11.3	0.24
proportion	-4853.8	144.7	65.3	20.0	0.22
max	-4873.2	164.1	65.3	18.1	0.22
base	-4874.7	165.6	65.5	18.2	0.22
overall	-5218.0	508.9	59.6	55.0	0.28
phraseology	-5528.4	819.3	57.8	59.6	0.24

Table 7: elpd_{loo} metrics for downstream application task (predicting vocabulary scores. Besides the main elpd_{loo}-metric, the table provides the difference to the elpd_{loo} to the best model, as well as the standard error for these two values (se and dse respectively).

The ELLIPSE dataset also provides other types of scores for student essays against which a comparison is possible. From those we selected the overall score, as it is the most important one, and the phraseology score, as it is the one closest related from vocabulary. By performing a regression with the same features on these scores, we can see whether the features are specific to vocabulary, as intended. Indeed, we find this to be the case for elpd_{loo} (see results in table 7 and figure 8),[28] despite the well-established halo effect, which leads annotators to provide roughly similar scores (e.g. Engelhard, 1994).

Although further research into the connection between content word replacement errors and vocabulary scores is required, the initial results show that our complexity scores can improve the performance of downstream applications.

gesting that the negative coefficient of the type-token ratio might be due to the essay length.

[28] It is not the case for R^2 in the case of the overall score, but this comparison is not directly admissible, because the variance for the Vocabulary scores (0.36) differs from that of the Overall score (0.41). The comparison of the elpd_{loo} is only acceptable because the number of data points and the predicted variables share a scale.

8 Conclusion

We proposes semantic error prediction as a task for investigating lexical semantic production complexity. Such an estimate of complexity is useful for many purposes in educational technology, including assessing output by learners and providing them with information for improving their writing skills.

Complex word identification systems, in contrast, are focused on difficulty in *comprehension* rather than *production*. Semantic error detection/correction system cannot be used this way, because they provide an estimate of how likely a word is to be wrong, not how difficult it was to produce the word in the first place. Semantic error prediction, thus, fills a gap in the CALL literature.

We propose and implement a method for creating semantic error prediction datasets from error correction datasets. Analysing the dataset with Bayesian logistic regressions, we found that verbs show a peculiar accumulation of semantic errors.

Furthermore, we train transformer-embedding based models for semantic error prediction. These models perform better than the baselines, although much room for improvement remains. Finally, we use the scores produced by the best of our models on the downstream task of predicting the vocabulary scores of student essays using a Bayesian linear regression. The results indicate that these lexical complexity scores improve the model.

Limitations

The present proposal suffers primarily from three limitations:

First, factors other than lexical semantic complexity might lead to content word replacement errors, rendering the proposed error prediction task an imperfect proxy. Future research should investigate other measures for active vocabulary for comparison.

Second, the error correction dataset used for our investigation does not provide information about important properties influencing error patterns, such as the first language of the L2 learners. However, our method is applicable to other datasets providing such information.

Third, our investigation is limited to an English error correction dataset. Error patterns might differ between languages. In some languages, for example morphologically richer languages, content word

replacement errors might be harder to identify or have a weaker connection to lexical semantics.

References

Mohammed Qassem Al-Shormani and Yehia Ahmed Al-Sohbani. 2012. Semantic errors committed by yemeni university learners: Classifications and sources. *International Journal of English Linguistics*, 2(66):p120.

Desislava Aleksandrova and Vincent Pouliot. 2023. Cefr-based contextual lexical complexity classifier in english and french. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, page 518–527, Toronto, Canada. Association for Computational Linguistics.

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*, page 79–88, New Orleans, Louisiana. Association for Computational Linguistics.

Aristotele. 1994. *Categories and De Interpretatione*, reprint edition. Clarendon Aristotle series. Clarendon Press, Oxford.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 Shared Task on Grammatical Error Correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, 49(3):643–701.

Bram Bulté and Alex Housen. 2012. *Defining and operationalising L2 complexity*, Language Learning & Language Teaching, page 21–46. John Benjamins Publishing Company.

Tomás Capretto, Camen Piho, Ravin Kumar, Jacob Westfall, Tal Yarkoni, and Osvaldo A. Martin. 2022. Bambi : A Simple Interface for Fitting Bayesian Linear Models in Python. *Journal of Statistical Software*, 103(15).

Council of Europe CoE. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment; Companion Volume*. Council of Europe Publishing, Strasbourg.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Anne Cutler. 1983. *Lexical complexity and sentence processing*, page 43–79. Wiley, Chichester, Sussex.

Vanessa De Wilde. 2023. Lexical characteristics of young l2 english learners' narrative writing at the start of formal instruction. *Journal of Second Language Writing*, 59:100960.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Julie L. Earles and Alan W. Kersten. 2017. Why are verbs so hard to remember? effects of semantic context on memory for verbs and nouns. *Cognitive Science*, 41(S4):780–807.

George Engelhard. 1994. Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of Educational Measurement*, 31(2):93–112.

May Fan. 2000. How big is the gap and how to narrow it? an investigation into the active and passive vocabulary knowledge of l2 learners. *RELC Journal*, 31(2):105–119.

Michael Flor, Steven Holtzman, Paul Deane, and Isaac Bejar. 2024. Mapping of american english vocabulary by grade levels. *ITL - International Journal of Applied Linguistics*.

Dedre Gentner and Ilene M. France. 1988. *The Verb Mutability Effect: Studies of the Combinatorial Semantics of Nouns and Verbs*, page 343–382. Morgan Kaufmann.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

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

Sylviane Granger. 1998. *The computer learner corpus: a versatile new source of data for SLA research*. Routledge.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Kipsamo E. Jeptarus and Patrick K. Ngene. 2016. Lexico-semantic errors of the learners of english: A survey of standard seven keiyo-speaking primary school pupils in keiyo district, kenya. *Journal of Education and Practice*, 7(13):42–54. ERIC Number: EJ1102824.

Nan Jiang. 2000. Lexical representation and development in a second language. *Applied Linguistics*, 21(1):47–77.

Mark D. Johnson. 2017. Cognitive task complexity and l2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing*, 37:13–38.

Alan W. Kersten and Julie L. Earles. 2004. Semantic context influences memory for verbs more than memory for nouns. *Memory & Cognition*, 32(2):198–211.

Aziz Khalil. 1985. Communicative error evaluation: Native speakers' evaluation and interpretation of written errors of arab efl learners. *TESOL Quarterly*, 19(2):335–351.

Minkyung Kim, Scott A. Crossley, and Kristopher Kyle. 2018. Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, 102(1):120–141.

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. 1998. The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, 19(2):255–271.

Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in l2 written production. *Applied Linguistics*, 16(3):307–322.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Philip M. McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.d., The University of Memphis, United States – Tennessee. 3199485.

Masashi Negishi, Tomoko Takada, and Yukio Tono. 2013. A progress report on the development of the cefr-j. In *Exploring language frameworks: Proceedings of the ALTE kraków conference*, page 135–163. Citation Key: negishi2013progress.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):179:1–179:42.

Musa Nushi, Roya Jafari, and Masoumeh Tayyebi. 2022. Iranian advanced efl learners' perceptions of the gravity of their peers' written lexical errors: The case of intelligibility and acceptability. *Iranian Journal of Foreign Language Teaching Innovations*, 1(1):41–56.

Margareta Olsson. 1972. *Intelligibility: A Study of Errors and Their Importance*. ERIC Number: ED072681.

Abril-Pla Oriol, Andreani Virgile, Carroll Colin, Dong Larry, Fonnesbeck Christopher J., Kochurov Maxim, Kumar Ravin, Lao Jupeng, Luhmann Christian C., Martin Osvaldo A., Osthege Michael, Vieira Ricardo, Wiecki Thomas, and Zinkov Robert. 2023. PyMC: A modern and comprehensive probabilistic programming framework in python. *PeerJ Computer Science*, 9:e1516.

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

Alice Pintard and Thomas François. 2020. Combining expert knowledge with frequency information to infer cefr levels for words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, page 85–92, Marseille, France. European Language Resources Association.

Muhammad Reza Qorib, Seung-Hoon Na, and Hwee Tou Ng. 2022. Frustratingly Easy System Combination for Grammatical Error Correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1964–1974, Seattle, United States. Association for Computational Linguistics.

Muhammad Reza Qorib and Hwee Tou Ng. 2023. System Combination via Quality Estimation for Grammatical Error Correction. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 12746–12759, Singapore. Association for Computational Linguistics.

Matthew Shardlow. 2013. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16, Tórshavn, Faroe Islands. LiU Electronic Press.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016. Swellex: Second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, page 76–84, Umeå, Sweden. LiU Electronic Press.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing Pretrained Language Models for Lexical Semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan. Asian Federation of Natural Language Processing.

A Further Results and Details

$$\textbf{B:} \quad \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + c$$

$$\textbf{B+I:} \quad \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + c + \beta_6 X_2 X_4$$

$X_1 =$ # characters

$X_2 =$ frequency

$X_3 =$ age of acquisition

$X_4 =$ word cefr-j level

$$X_5 = \begin{cases} 1 & \text{if token is a verb} \\ 0 & \text{in other case} \end{cases}$$

$c = (\beta_{A2} C_{A2} + \beta_{B1} C_{B1} + \ldots) =$ effect of student CEFR level

Figure 6: Equations describing the two Bayesian logistic regression models: Basic (B) and Basic with Interaction added (B+I).

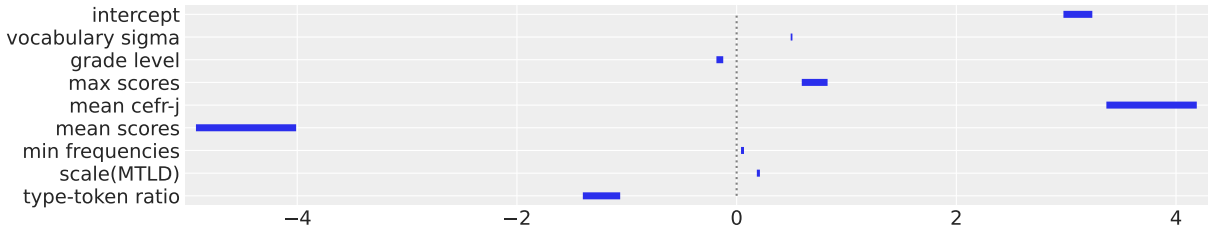


Figure 7: HDIs for Max+Mean model predicting the vocabulary scores. Max and mean scores refer to the pooled results of our neural model.

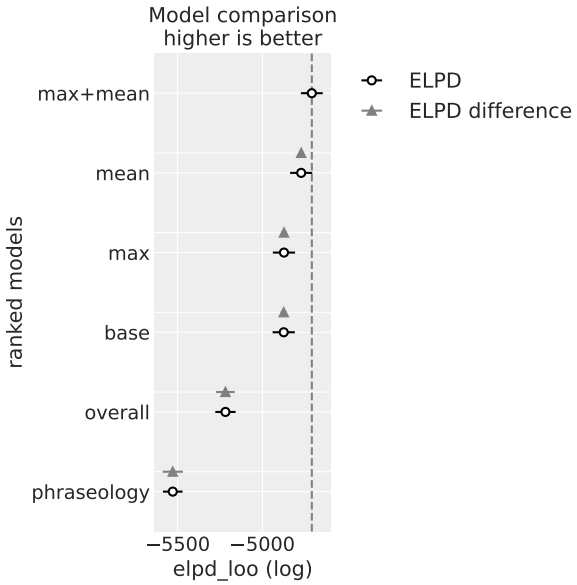


Figure 8: elpd_{loo} scores for Bayesian linear regression models predicting vocabulary scores (top 4 model) as well as Overall scores and Phraseology scores.

	space	BERT	RoBERTa	LLAMA2
midrule batch size	{640, 1280, 1920, 2560, 3200}	2560	2560	640
learning rate	$\{1 \cdot 10^{-2}, 5 \cdot 10^{-3}, 1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 5 \cdot 10^{-5}, 1 \cdot 10^{-5}\}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-5}$
epochs	{20, 30, 40, 50, 60}	50	50	40

Table 8: Hyperparameters search space and selected hyperparameters.