Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

LLM-Generated Contexts to Practice ESP Vocabulary: Corpus Presentation and Comparison

Iglika Nikolova-Stoupak[1] Serge Bibauw[3] Amandine Dumont[2] Françoise Stas[2]
Patrick Watrin[1] Thomas François[1]
(1) CENTAL, (2) ILV, (3) IACCHOS, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgique
{iglika.nikolova, serge.bibauw, amandine.dumont, francoise.stas,
patrick.watrin, thomas.francois}@uclouvain.be

RÉSUMÉ

Contextes générés par LLM pour pratiquer le vocabulaire anglais de spécialité : présentation de corpus et comparaison

Ce projet analyse la capacité des LLM (grands modèles de langue) et de corpus web à fournir des contextes visant la pratique et l'apprentissage de vocabulaire anglais spécialisé dans un contexte universitaire. Le but sous-jacent est d'augmenter le volume d'exemples utilisables et leur facilité de mise au point tout en conservant la qualité actuelle où ils sont conçus par des spécialistes. Sur la base d'un jeu de contextes de référence — utilisés en classe — visant à l'apprentissage d'une liste de vocabulaire spécialisé, nous comparons les caractéristiques linguistiques de contextes générés par trois LLM récents de différentes tailles (Mistral-7B-Instruct, Vicuna-13B et Gemini 1.0 Pro) et un corpus de contextes extraits automatiquement d'articles de sites web spécialisés. Les caractéristiques textuelles évaluées incluent la longueur, la morphosyntaxe, la sémantique et le niveau discursif. En fin de compte, nous identifions le corpus généré par un LLM (Gemini) dans un scénario one-shot comme étant celui qui se rapproche le plus du corpus de référence.

ABSTRACT

This project analyses the ability of LLMs (large language models) and web-based corpora to provide contexts for the practice and acquisition of specialised English vocabulary in a university context. The underlying purpose is to increase the volume of usable examples and their ease of generation while retaining the currently established quality of learning materials as crafted by specialists. We present a reference corpus of contexts — handpicked by expert teachers — for a specialised vocabulary list, as well as related corpora generated by three recent LLMs of different sizes (Mistral-7B-Instruct, Vicuna-13B, and Gemini 1.0 Pro) and a corpus extracted from articles crawled from specialised websites. We evaluate and compare the corpora based on a representative set of textual characteristics (length-based, morphosyntactic, lexico-semantic, and discourse-related). Ultimately, we identify a corpus generated by an LLM (Gemini) in a one-shot setting as coming closest to the reference one.

MOTS-CLÉS : grands modèles de langue, vocabulaire anglais spécialisé, traits de lisibilité.

KEYWORDS: large language models, specialised English vocabulary, readability features.

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 472–498.

1 Introduction

The present study is part of a broader project conducted at Université catholique de Louvain that aims at leveraging natural language processing (NLP) tools to facilitate the acquisition of English for specific purposes (ESP) vocabulary by providing multiple examples of its natural use in context. The project involves several university courses mapped to proficiency levels B1 to C1 and designed to teach ESP to STEM (science, technology, engineering, and mathematics) students. In these courses, students have to master predefined vocabulary lists comprised of both strictly scientific vocabulary (e.g. "a chemical") and other vocabulary that commonly appears in scientific contexts (e.g. "to assume"). A preliminary survey revealed that the target students currently study the lists as is, at best using flash cards. Research has, however, shown the importance of regularly exposing learners to words in authentic and informative contexts (Huckin & Coady, 1999; Ramos & Dario, 2015; Godwin-Jones, 2018). Collecting authentic contexts is, unfortunately, very time-consuming for ESP teachers, which generally impedes offering enough authentic contexts to students as support for vocabulary learning. This is why we intend to automatically retrieve and generate contexts of use for any target ESP word, thereby significantly decreasing the burden in terms of both time and effort that currently results from the manual collection of context sentences.

In this study, we explore two ways of collecting such contexts : firstly, their extraction from a large corpus made up of websites routinely exploited by ESP teachers ; and secondly, their generation using three large language models (LLMs) : Mistral-7B-Instruct, Vicuna-13B, and Gemini 1.0 Pro. More specifically, this paper analyses the linguistic characteristics – using standard readability- and stylometry-related variables – of sentences in our web-crawled and LLM-generated corpora and their similarities and differences with a reference corpus. The last is made up of examples handpicked by ESP teachers among vocabulary test materials used in university ESP courses.

2 Background

2.1 Automatic Text Retrieval/Generation in Language Learning

The advancement of the Internet and Big Data has long been viewed as an opportunity for language teachers and learners to get hold of a large quantity of learning materials that are characterised with authenticity and timeliness. One of the established roles of NLP in EFL studies is the retrieval of relevant materials from the web, often followed by their evaluation, annotation and/or adaptation and the generation of related exercises (Litman, 2016; Meurers, 2021). Via a survey, Wilson (2004) evaluates students' practices and satisfaction in relation to the use of web resources for independent ESL study as well as assembles a list of recommended websites for learners. Heilman *et al.* (2008) first create a corpus of web-crawled texts to be used for vocabulary and reading practice and then develop a system called REAP Search that allows the selection of particular texts from the corpus based on defined constraints (e.g. the presence of specific words). Similarly, Yoon *et al.* (2017) retrieve a number of YouTube videos based on criteria such as the existence of manual transcriptions and go on to use them in the generation of listening exercises for the TOEIC certificate exam. Other studies (Meurers *et al.*, 2010; Hussin *et al.*, 2010; Jin & Lu, 2018) focus less on the specificities of web-based textual sources than on their later enrichment and annotation for use in language learning.

A recent and revolutionary technology capable of producing humanlike language, LLMs have already

been exploited in a variety of scenarios, including the creation of EFL learning materials. Young & Shishido (2023a) had ChatGPT produce text of different proficiency levels based on articles from an online newspaper. The levels' correctness was confirmed through a readability analysis. In another experiment, Young & Shishido (2023b) used ChatGPT for the generation of dialogues. The general topic and participants were indicated in the prompts, and multiple readability metrics were used to analyse the suitability of the derived dialogues and to determine their best target audience. They concluded that the dialogues are most suitable for the A2 level (followed by B1), while students of higher proficiency levels may miss out on elements like colloquial expressions and phrasal verbs. Shaikh *et al.* (2023) made use of a questionnaire to evaluate users' views on the effectiveness of ChatGPT in specialised EFL studies. Students of different nationalities, proficiency levels and fields of study were asked to converse with ChatGPT on different topics, and engage in vocabulary practice and have the virtual assistant edit text they produce. Students' opinions were generally favourable, in particular with regard to ChatGPT's assistance in vocabulary acquisition.

2.2 Readability Features

Readability, often considered to date back to Sherman's experiments in 1893, is a primary measure used for quantitative description of text (DuBay, 2007). Its main purpose is the estimation of a textual unit's reading difficulty and, thereby, appropriateness for a given audience (typically, children of a certain age). To this aim, numerous readability formulas have been developed throughout the years [1], relying on a variety of shallow textual characteristics (such as sentence or word length), more advanced ones (e.g. syntax or discourse properties), or comparison against vocabulary lists. Although state-of-the-art readability estimations are now mostly provided by deep neural models (Vajjala, 2022), readability formulas based on engineered features have dominated the field for almost 90 years, and some of the best-performing current systems rely on both deep learning and such features within a hybrid architecture (Deutsch *et al.*, 2020; Wilkens *et al.*, 2024).

Features are central to readability because they link the theory of the reading processes to the pragmatic approach typical to predictive modelling. The reading process is made up of three main steps : visual perception, decoding, and comprehension, and each of them can be impacted by a given text's characteristics (François, 2011). For instance, more frequent words are generally decoded faster than rare ones, some syntactic structures seem harder to parse for the brain than others, and lexemes representing abstract concepts are generally activated more slowly in the brain than more concrete ones. There is a large amount of psycholinguistic studies that have stressed a specific aspect of language that is likely to impact the reading process (Ferrand, 2007). In this work, we exploit the long tradition of readability variables, hundreds of which have been investigated, parametrised and tested on different corpora since the 1920s.

Many of the mentioned features have also been utilised in textual descriptions that are not strictly related to complexity. The field of second language acquisition, in particular, aims to describe the language produced by language learners and also resorts to various features, some of which overlap with readability ones. For instance, "lexical richness" (close to "lexical diversity"), strongly interconnected with type-to-token ratio, is a concept defined by Yule (1944) and used to estimate a particular author's (or text's) distinctive linguistic characteristics. A related term used in stylometry is "lexical sophistication", typically associated with word frequency and concreteness (Kyle, 2019). Other metrics aim to measure "lexical density", which refers to the proportion of content words in the

1. For surveys of the field, see François (2011); Collins-Thompson (2014); Vajjala (2022).

text (Ure, 1971). Cech & Kubat (2018) specifically refer to the "morphological richness" of a text, defined as the difference between the vocabulary richness (i.e. type-to-token ratio) of lemmas and words, as an important characteristic in authorship attribution.

3 Methods

In the current study, we collect and analyse two general types of contexts for ESP vocabulary learning : generated by LLMs (Mistral-Instruct, Vicuna, and Gemini Pro) and obtained through web-crawling. Figure 1 illustrates the different steps of this procedure. The resulting corpora, described in Section 3.1, are compared to a reference ESP corpus associated with the same vocabulary items, handpicked by teachers on the basis of a set of stylistic features introduced in Section 3.2. We hypothesise that the most adequate generation method produces contexts closest to the reference corpus in terms of stylistic characteristics.

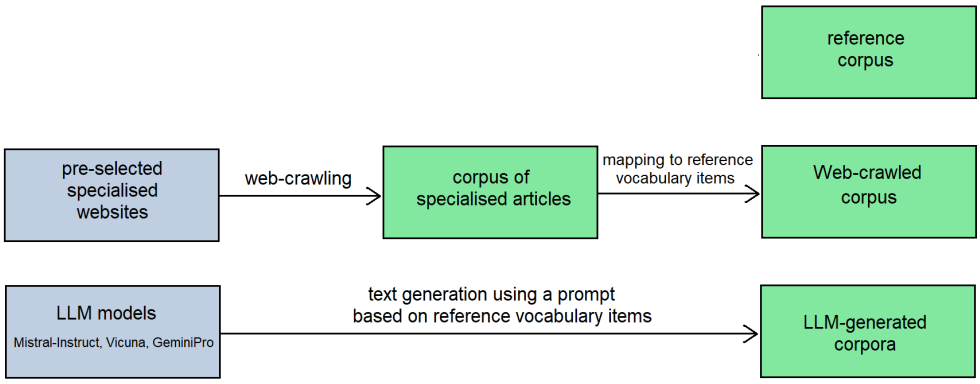

FIGURE 1 – Collection procedure for the examined corpora

3.1 Derivation of the Corpora

3.1.1 Reference

The reference corpus, consisting of 244 contexts, is crafted by ESP teachers from Université catholique de Louvain and consists of a sample of study and examination materials used in the acquisition of vocabulary knowledge in ESP courses. The provided contexts are typically one to three sentences long and reflect key pedagogical qualities as defined by the ESP teachers, including appropriate length, topic relevance, authenticity, timeliness and level-appropriate vocabulary and grammar. The corpus's items are further classified as belonging to CEFR levels B1 or B2 and to the fields of "agronomy" or "general science".

3.1.2 Web-Crawled

The web-crawled context corpus is extracted from a large corpus of articles found in 36 websites (centred around the domains of agronomy, civil engineering and general science), which are commonly

used by ESP teachers when they manually craft context examples [2]. The articles and associated metadata [3] were extracted using the Python tools *beautifulsoup4* [4] and *newspaper* [5] and then preprocessed for noise removal. For the current experiment, sentences in the articles were mapped to the vocabulary items associated with the reference corpus. The recorded metadata was used to ensure compatibility of domains. A set of heuristic rules defined the search for joint matches in terms of words and their POS tags, the latter allowing for lemma rather than word correspondence and resolving cases of polysemy that imply words' parts of speech (e.g. "yield" - both noun and verb). The quality of the issuing corpus was verified manually, and issues that can be fixed automatically (mostly format-related) were identified and resolved. Quality problems that are not readily fixable included text-conversion errors (4 contexts featured mistakes of this type), sentence fragments (a total of 3), and a run-on sentence. Levenstein distance was applied to ensure that no sentences are (closely) identical.

3.1.3 LLM-Generated

To generate contexts based on a prompt, three recent, easily accessible LLMs of different sizes were selected : Mistral-Instruct (7B parameters), Vicuna (13B) and Gemini 1.0 Pro (600B).

Mistral, developed by Mistral AI, makes use of grouped-query and sliding window attention mechanisms, thus substantially increasing inference speed and reducing memory constraints at decoding. Its finetuned "Instruct" version demonstrates superior performance than LLaMA on human as well as automated benchmarks (Jiang *et al.*, 2023). Following a process of trial and error, the prompt's role was set as "system" rather than "user", thus placing focus on the text generation guidelines.

The Vicuna model is based on LLaMA as enhanced via instruction-tuning on data provided by ShareGPT, a platform that features full conversations of users with ChatGPT and facilitates their sharing and reuse (Mehta, 2022). It thus makes use of ChatGPT's established linguistic abilities. Vicuna is associated with better privacy as compared with ChatGPT and can reach as much as 90% of the latter's performance despite its compact size (Lam *et al.*, 2023). In this study, both the Mistral and Vicuna models were used through the "LM studio" interface.

Gemini is a state-of-the-art multimodal model released by Google DeepMind in three versions : Ultra, Pro and Nano. It uses advanced attention mechanisms, such as multi-query attention, and supports a context length of 32k. Its Ultra version achieves higher performance than GPT-4 in 30 out of 32 language benchmarks and uniquely surpasses human performance on the exam benchmark MMLU (Anil *et al.*, 2023). Trained for increased deployability, the Pro version almost matches GPT-3.5 in performance (Akter *et al.*, 2023). Even though the model is proprietary, at the time of writing, it can be accessed freely via the Google AI Studio developer tool within a given quota.

A prompt format was defined and tested that includes the vocabulary item's associated domain and CEFR proficiency level, the part of speech in the case of nouns, verbs, adjectives and adverbs and differentiation between word and expression [6]. Mistral output demonstrated high variance based on its temperature setting [7]. A value of 0.8 was opted for as the threshold below which output was perceivably too homogeneous and consisted of definitions of the target vocabulary rather than

2. See Appendix 1 : List of Crawled Websites for the list of utilised websites
3. title, date, scientific domain, format (html vs pdf)
4. Version 4.12.3 ; https ://pypi.org/project/beautifulsoup4/
5. Version 0.2.8 ; https ://pypi.org/project/newspaper3k/
6. See Appendix 2 : Prompts used for LLM Generation for the utilised prompts
7. a model's temperature defines its level of unexpectedness or creativity

examples of use.

An additional experiment was carried out using one of the models (Gemini was opted for due to its fastest performance), namely a one-shot setting in which we offer the corresponding example from the reference corpus. The purpose was to test the LLM's efficiency in adapting its output to the given example and the underlying potential for multiple contexts per target word to be derived from a single professionally-crafted one.

The models were asked to provide output until it was automatically verified that output was present and that it contained the respective vocabulary items (verbatim or, in the case of verbs and nouns, in any possible form). Readily fixable issues (such as an additional "Explanation" part) were addressed manually, and no more substantial issue was found. The Gemini model exhibited by far the fastest performance (650 seconds, compared with 2624 for Mistral-Instruct and 5447 for Vicuna). Once again, it was ensured that no sentences were (closely) identical [8].

3.2 Stylistic Comparison

For the stylistic analysis of the various contexts, we selected various atomic features related to textual readability belonging to four general categories : length-based, morphosyntactic, lexico-semantic, and discourse-related [9].

The selected length-based features are the numbers of words and syllables per sentence and numbers of letters and syllables per word. Morphosyntactic features include the number of noun phrases per sentence, the number of non-stem words per sentence, the number of punctuation signs per sentence (excluding end-of-sentence punctuation), the percentage of sentences ending in question and exclamation marks, and the overall morphological richness as defined by Cech & Kubat (2018). The selected lexico-semantic features are the number of verbs, first-person pronouns, proper nouns and the joint number of adjectives and adverbs per sentence. We also considered the word-based and lemma-based type-to-token ratios, the percentage of hapax legomena, the percentage of words not present in the Dale-Chall list, the average concreteness (based on Brysbaert *et al.* (2014)'s rated concreteness list of 40k English lemmas), and the 10 most frequent words including and excluding stop words. Finally, the discourse-related features consist of the number of pronouns per sentence, the percentage of anaphora-denoting words per sentence [10], and the cosine distance between all sentences in the respective corpus [11]. The assignment of features to a particular category is occasionally highly subjective ; for instance, the number of verbs in a sentence could be interpreted as being more strongly related to a sentence's syntactic structure than its semantics.

In their work on readability classification, Wilkens *et al.* (2022) concluded that the use of a set of aggregators provides a better estimation of textual qualities than a single selected value. Where relevant, the average, minimal and maximal values for a feature, as well as the standard deviation (SD), were examined. This allowed for comparisons of both the texts' general qualities (as often relevant to a measure of complexity) and the span of values contained (which can serve as an estimation of

8. Our experimental setup featured an 11th Gen Intel Core i7 CPU with 8 cores and TigerLake-LP GT2 integrated GPU.
9. Measures that strongly imply a larger textual unit (e.g. textual cohesion) were naturally excluded.
10. The words considered are the following : definite article (*the*) ; personal pronouns (*he*, *she*, *it*, *they*) ; demonstrative pronouns (*this*, *that*, *these*, *those*) ; relative pronouns (*who*, *which*, *whose*, *whom*, *where*) ; indefinite pronouns (*all*, *some*, *none*, *any*, *each*, *every*) ; adverbs (*here*, *there*, *now*, *then*)
11. calculated using Python's *transformers* library ; sentence transformer model *paraphrase-MiniLM-L6-v2*

textual variety). [12]

The following steps were taken to evaluate continuous features. Firstly, a Shapiro-Wilk test (Shapiro & Wilk, 1965) was used to determine whether the features demonstrate normal distribution. As the only feature that was normally distributed was the percentage of non-stem words per sentence, we used Mann-Whitney U, a non-parametric test, to determine whether differences between the reference corpus and the rest of the corpora were significant. Statistical significance, when present, was assigned one of three levels corresponding to p-values of 0.001, 0.01 and 0.05.

4 Results

The results of our stylistic comparison are reported in Table 1. Only the most relevant features have been listed ; please refer to Appendix 4 : Detailed Results for a comparison of all features. This section provides an overview of the main results, organised according to the four families of features.

Feature	Ref.	Web	Mistral	Vicuna	Gemini	Gemini : one-shot
words in sample	9823	7269	4615	4852	4160	**10366**
words / sentence	*13.59*	*14.93****	*9.34***	*9.6***	*8.51****	***12.26******
letters / word	*5.29*	*5.38*	*5.3*	*5.43*	*5.6**	***5.53***
noun phrases / sentence	*5.87*	*8.16****	*5.56*	*5.53*	*4.92****	***5.39****
non-stem words / s-ce	*33.56*	*31.56****	*35.14****	*35.59****	*36.36****	***36******
punctuation signs / s-ce	*1.56*	*2.7*	*0.77****	*0.99****	*0.75****	***1.25****
verbs / sentence	*2.45*	*3.83*	*2.54****	*2.44****	*2.27****	***2.53***
adj. and adv. / sentence	*2.96*	*4.13****	*2.21****	*2.31****	*2.29****	***2.52*****
1st-person pron. / s-ce	*0.1*	*0.12*	*0.39****	*0.11*	*0.06***	***0.07****
proper nouns / sentence	*0.9*	*1.46*	*0.06****	*0.32****	*0.1****	***0.27******
hapax legomena	16.13	22.56	16.25	18.18	19.75	**14.14**
concreteness	2.46	2.36	2.44	2.44	2.42	**2.41**
pronouns / sentence	*0.88*	*1.27*	*1.06*	*0.8*	*0.5****	***0.73****
anaphora words / s-ce	*20.49*	*10.78*	*10.47*	*10.43*	*13.42****	***24.59***
cos. distance btwn s-ces	*0.14*	*0.1****	*0.18****	*0.17****	*0.18****	***0.15*****

TABLE 1 – Comparison of the corpora based on a sample of textual features. The average values of continuous characteristics are indicated in *italics*, and the statistical significance of their divergence from the reference corpus is marked with * (lowest), ** and *** (highest). The one-shot Gemini corpus is represented in **bold** to denote its highest global closeness to the reference as per Section 4.5.

4.1 Length-Based Features

The one-shot corpus contains the largest number of words, closely followed by the reference and web-crawled ones and then by the three zero-shot LLM-generated corpora, which exhibit similar values. The number of words per sentence, tightly associated with textual complexity, is highest in the web-crawled corpus, followed by the reference one and then by all LLM corpora. Differences at the

12. Refer to Appendix 3 : Features Used in Corpus Comparison for an overview of all investigated features.

"word" level (e.g. the number of letters per word) are minimal. The ranges of length-based features [13] are lowest with the LLM corpora (implying a lack of variety) and highest with the web-crawled one, followed closely by the reference corpus. The web-crawled corpus tends to demonstrate the largest SD. In this category, the Vicuna and Mistral corpora demonstrate the lowest significance in difference with the reference.

4.2 Morphosyntactic Features

The number of noun phrases per sentence is similar between the reference corpus and the LLM ones and significantly higher within the web-crawled corpus. The number of non-stem words, which is the most statistically different feature in the category, is lowest in the web-crawled corpus (interestingly suggesting lower complexity) and highest in the LLM-generated ones, the reference corpus standing in the middle. In relation to the number of punctuation signs, the reference corpus is once again in the middle, this time the web-crawled corpus exhibiting the highest value. The LLM corpora do not demonstrate variety in end-of-sentence punctuation. Morphological richness is stable at 0.02 for all corpora. Once again, the web-crawled corpus has the highest SD and value ranges are typically narrower for LLM corpora.

4.3 Lexico-Semantic Features

The number of verbs per sentence (associated with the presence of complex sentences) is closely stable, with the exception of the web-crawled corpus, where it is significantly higher. The number of adjectives and adverbs (which demonstrates the highest statistical difference in the category), as well as the number of proper nouns per sentence, are highest within the web-crawled corpus and lowest within the LLM ones, the reference corpus standing in the middle. The most hapax legomena are found in the web-crawled corpus and the fewest in the LLM ones (in particular, the one-shot corpus), implying re-use of vocabulary. First-person pronouns are generally rare, the highest value of 0.39 per sentence being associated with the Mistral corpus. The percentage of words outside of the Dale-Chall frequency list is highly stable, and so are the average concreteness of words and type-to-token ratios. Once again, the web-crawled corpus demonstrates the highest value ranges and SD. With the exception of the web-crawled corpus, the most frequent words excluding stop words are narrowly related to the texts's specialisation (e.g. "water", "climate"). When stop words are retained, the words are highly identical, the Mistral corpus uniquely featuring the pronoun "I".

4.4 Discourse-Related Features

Cosine distance (i.e. the estimated semantic difference between examples) diverges the most from the reference corpus. While average cosine distance demonstrates similar values, the maximal one is highest with the LLM corpora. The number of pronouns per sentence is highest in the web-crawled corpus and varies among the LLM ones. Vicuna's average value is closest to the reference but at the expense of a significant difference in distribution. Anaphora-denoting words are most prominent in the one-shot followed by the reference corpus, values being significantly lower in the other corpora. Uniquely for this category, the web-crawled corpus does not exhibit high SD values.

13. i.e. the differences between their maximal and minimal values

4.5 Additional experiments

In this subsection, we report three additional experiments that we carried out for the purpose of gaining deeper insight about our corpora.

First, we divided each corpus into two according to the proficiency level of the target vocabulary items (B1 or B2). In such a scenario, the web-crawled corpus, which consists of authentic texts not specifically conceived for language learners, increases in significance of difference with the reference one ; in particular, in relation to level B2, for which it exhibits 12 significantly different features (as opposed to 5 when the entire corpus is considered).

In contrast, sensitivity in relation to proficiency levels is noticeable among the LLM corpora. Features associated with textual complexity, such as the total number of words, the number of letters per word and the number of first-person pronouns per sentence, demonstrate significantly differing values compared to when the corpus is taken in its entirety. In particular, the Gemini corpus shows the highest modification of values based on proficiency level. Interestingly, there are even cases where LLM corpora show higher sensitivity to the level at hand than the reference corpus [14]. As the different CEFR levels are also associated with different domains, the most frequent words in the LLM corpora now reflect the domain at hand (derivatives of "science" being common for the scientific domain and words such as "crop" and "soil" for agronomy).

As a second experiment, the complete set of textual characteristics was used within a global distance metric in order to determine which corpus is globally closest to the reference one. For this purpose, min-max normalisation was applied, and the Euclidean distance between corpora was calculated. The one-shot Gemini corpus ranked first (2.96), followed by the web-crawled one (3.8) and the three zero-shot LLM corpora, which in turn demonstrated relatively similar values [15].

Finally, as regards the one-shot scenario, the features of the corresponding corpus were compared to those of its zero-shot counterpart, revealing that their majority [16] come closer to the reference, reducing the significance in divergence in 10 out of all 13 cases. The one-shot generation method also leads to significantly increased global closeness to the reference.

5 Discussion

Given the high value ranges and SD pertaining to the web-crawled corpus, as well as the tendency between its and LLM-generated texts to diverge from the baseline in opposite directions, the two types of corpora can work together effectively within an educational framework to provide a variety of contexts for ESP vocabulary. The web-crawling method is computationally efficient and tends to provide texts that are naturally close to the reference examples. In turn, the benefits of LLM generation include sensitivity to the CEFR level at hand and high malleability, including the potential to derive a large number of examples from a single one within a one-shot setting.

Table 2 shows the example sentences for the verb "to avoid" in all discussed corpora. In accordance

14. For instance, when solely B1 examples are considered, the reference corpus unintuitively has higher values for the percentages of non-stem words and words outside the Dale-Chall frequency list compared to when the entire corpus is considered.

15. Mistral : 4.03 ; Vicuna : 4.06 ; Gemini : 4.57

16. excluding "percentage of hapax legomena", "average concreteness" and several standard deviation values

with the results stated in Section 4, the web-crawled example exhibits high complexity : it is longest and contains three proper nouns and a direct quotation. The Vicuna and Gemini examples show similarity with the reference if one considers the number of verbs and the joint number of adjectives and adverbs [17]. Mistral adds variety with its use of first-person language, which however results in the example's reduced formality and in-domain quality. In turn, Gemini's example is narrowly associated with the scientific domain. Much akin to the reference, the "Gemini : one-shot" example features additional vocabulary items that are relevant for the same learner audience ("essential", "biases", "skew"), yet it does not demonstrate any perceivable copying of the reference content.

Reference	There is still time to reverse the warming trend and **avoid** global environmental and economic catastrophe.
Web-Crawled	"There are few institutional structures to achieve co-operation globally on the sort of scales now essential to **avoid** very serious consequences," warns lead author Dr Brian Walker of Australia's CSIRO.
Mistral	I try to **avoid** using my phone while I study because it can be a great distraction.
Vicuna	To **avoid** overfishing, it is essential to manage fisheries sustainably and establish marine protected areas.
Gemini	To **avoid** contamination, the scientist carefully wore gloves and a lab coat while conducting the experiment.
Gemini : one-shot	In scientific experiments, it is essential to **avoid** biases that could skew the results.

TABLE 2 – Examples for the vocabulary item "avoid" (domain "science" ; level "B1")

6 Conclusion and Future Directions

This work discussed context corpora derived via two discrete NLP-based methods (web-crawling and generation by LLMs), which can be used to aid university students in the acquisition of ESP vocabulary. The stylistic characteristics of the generated contexts were compared to a professionally crafted baseline of context examples through quantitative evaluation based on readability features. The comparison revealed higher similarity among different LLM-generated corpora than between them and corpora of a different nature. The "Gemini : one-shot" corpus was discovered to be globally closest to the reference, thereby showing that generation can be refined through prompt engineering. Future experiments including few-shot use of only several high-quality examples independent of a vocabulary list can help mitigate the current limitation of reliance on a reference corpus.

Associated future plans include an evaluation of the pedagogic characteristics of contexts as opposed to their readability-based qualities. In more practical terms, pre- and post-tests of students' performance will be conducted in relation to the introduction of a large inventory of NLP-derived examples in the implied university courses.

17. respectively, 3 and 3 for the reference, 4 and 4 for Vicuna and 3 and 1 for Gemini

References

AKTER S. N., YU Z., MUHAMED A., OU T., BÄUERLE A., CABRERA A. A., DHOLAKIA K., XIONG C. & NEUBIG G. (2023). An in-depth look at Gemini's language abilities. arXiv : 2312.11444.

ANIL R., BORGEAUD S., WU Y., ALAYRAC J.-B., YU J., SORICUT R., SCHALKWYK J., DAI A. M., HAUTH A., MILLICAN K., SILVER D., PETROV S., JOHNSON M., ANTONOGLOU I., SCHRITTWIESER J., GLAESE A., CHEN J., PITLER E. & VINYALS O. (2023). Gemini : A family of highly capable multimodal models. arXiv : 2312.11805.

BRYSBAERT M., WARRINER A. B. & KUPERMAN V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, **46**(3), 904–911. DOI : 10.3758/s13428-013-0403-5.

CECH R. & KUBAT M. (2018). Morphological richness of text. In M. FIDLER & V. CVRCEK, Éds., *Taming the Corpus*, p. 63–77. Springer. DOI : 10.1007/978-3-319-98017-1_4.

COLLINS-THOMPSON K. (2014). Computational assessment of text readability : A survey of current and future research. *International Journal of Applied Linguistics*, **165**(2), 97–135.

DEUTSCH T., JASBI M. & SHIEBER S. (2020). Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* : Association for Computational Linguistics. DOI : 10.18653/v1/2020.bea-1.1.

DUBAY W. H. (2007). *The Classic Readability Studies*. Rapport interne, ERIC Clearinghouse. DOI : 10.1109/TPC.2008.2007872.

FERRAND L. (2007). *Psychologie cognitive de la lecture*. Bruxelles : De Boeck.

FRANÇOIS T. (2011). La lisibilité computationnelle : un renouveau pour la lisibilité du français langue première et seconde ? *International Journal of Applied Linguistics (ITL)*, **160**, 75–99.

GODWIN-JONES R. (2018). Evolving views on vocabulary development. *Language Learning & Technology*, **22**(3), 1–19.

HEILMAN M., ZHAO L., PINO J. & ESKENAZI M. (2008). Retrieval of reading materials for vocabulary and reading practice. In J. TETREAULT, J. BURSTEIN & R. DE FELICE, Éds., *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, p. 80–88, Columbus, Ohio : Association for Computational Linguistics.

HUCKIN T. & COADY J. (1999). Incidental vocabulary acquisition in a second language : A review. *Studies in second language acquisition*, **21**(2), 181–193.

HUSSIN A., CHAN Y. F. & ZUBAIDAH A. (2010). Scientific structural changes within texts of adapted reading materials. *English Language Teaching*, **3**. DOI : 10.5539/elt.v3n4p216.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L. & EL SAYED W. (2023). Mistral 7B. arXiv : 2310.06825.

JIN T. & LU X. (2018). A data-driven approach to text adaptation in teaching material preparation : Design, implementation, and teacher professional development. *TESOL Quarterly*, **52**, 457–467. DOI : 10.1002/tesq.434.

KYLE K. (2019). Measuring lexical richness. In S. WEBB, Éd., *The Routledge Handbook of Vocabulary Studies*, p. 454–476. Routledge. DOI : 10.4324/9780429291586.

LAM K.-Y., CHENG V. C. W. & YEONG Z. K. (2023). Applying large language models for enhancing contract drafting. In *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace*, p. 70–80.

LITMAN D. J. (2016). Natural language processing for enhancing teaching and learning. In *AAAI Conference on Artificial Intelligence*.

MEHTA I. (2022). ShareGPT lets you easily share your ChatGPT conversations. Blog post.

MEURERS D. (2021). *Natural Language Processing and Language Learning*, In *The Encyclopedia of Applied Linguistics*, p. 1–15. John Wiley Sons, Ltd. DOI : https ://doi.org/10.1002/9781405198431.wbeal0858.pub2.

MEURERS D., ZIAI R., AMARAL L., BOYD A., DIMITROV A., METCALF V. & OTT N. (2010). Enhancing authentic web pages for language learners. In J. TETREAULT, J. BURSTEIN & C. LEACOCK, Éds., *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 10–18, Los Angeles, California : Association for Computational Linguistics.

RAMOS R. & DARIO F. (2015). Incidental vocabulary learning in second language acquisition : A literature review. *Profile Issues in TeachersProfessional Development*, **17**(1), 157–166.

SHAIKH S., YAYILGAN S. Y., KLIMOVA B. & PIKHART M. (2023). Assessing the usability of ChatGPT for formal english language learning. *European Journal of Investigative Health Psychology and Education*, **13**, 1937–1960. DOI : 10.3390/ejihpe13090140.

SHAPIRO S. S. & WILK M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**(3/4), 591–611. DOI : 10.2307/2333709.

URE J. (1971). Lexical density and register differentiation. *Applications of linguistics*, **23**(7), 443–452.

VAJJALA S. (2022). Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 5366–5377.

WILKENS R., ALFTER D., WANG X., PINTARD A., TACK A., YANCEY K. P. & FRANÇOIS T. (2022). FABRA : French aggregator-based readability assessment toolkit. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1217–1233, Marseille, France : European Language Resources Association.

WILKENS R., WATRIN P., CARDON R., PINTARD A., GRIBOMONT I. & FRANÇOIS T. (2024). Exploring hybrid approaches to readability : experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics : EACL 2024*, p. 2316–2331.

WILSON R. (2004). Computers and the internet : Together a great tool for esl/efl learners.

YOON S.-Y., LEE C. M., HOUGHTON P., LOPEZ M., SAKANO J., LOUKINA A., KROVETZ B., LU C. & MADNANI N. (2017). Analyzing item generation with natural language processing tools for the toeic ® listening test : Analyzing item generation with nlp tools. *ETS Research Report Series*, **2017**. DOI : 10.1002/ets2.12183.

YOUNG J. C. & SHISHIDO M. (2023a). Evaluation of the potential usage of ChatGPT for providing easier reading materials for ESL students. In *Proceedings of EdMedia and Innovate Learning*, p. 155–162 : Association for the Advancement of Computing in Education (AACE).

YOUNG J. C. & SHISHIDO M. (2023b). Investigating OpenAI's ChatGPT potentials in generating chatbot's dialogue for English as a foreign language learning. *International Journal of Advanced Computer Science and Applications*, **14**(6), West Yorkshire. DOI : 10.14569/IJACSA.2023.0140607.

YULE G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Appendix 1 : List of Crawled Websites

https://climate.ec.europa.eu/climate-change_en
https://climate.ec.europa.eu/eu-action_en
https://climate.ec.europa.eu/index_en
https://climate.nasa.gov/
https://engineeringdiscoveries.com/
https://newatlas.com
https://sciencedemonstrations.fas.harvard.edu/
https://sustainability.stanford.edu/
https://world-nuclear.org
https://www.advancedsciencenews.com
https://www.computerworld.com/
https://www.eurekalert.org/
https://www.green.earth/
https://www.iea.org/
https://www.ipcc.ch
https://www.livescience.com/
https://www.nationalgeographic.org/society/
https://www.nature.com/
https://www.ncbi.nlm.nih.gov/
https://www.networkworld.com/
https://www.newscientist.com/
https://www.npr.org/sections/science/
https://www.pcworld.com
https://www.pewresearch.org/topic/internet-technology/
https://www.pewresearch.org/topic/science/
https://www.popularmechanics.com/
https://www.science.org/
https://www.sciencealert.com/
https://www.sciencedaily.com/
https://www.scienceopen.com/
https://www.scientificamerican.com
https://www.triplepundit.com/
https://www.un.org/en/
https://www.un.org/en/climatechange
https://www.usgs.gov/programs/earthquake-hazards/
https://www.wwf.org.uk

Appendix 2 : Prompts used for LLM Generation

Zero-shot setting for Mistral, Vicuna, and Gemini :

> Here is a sentence[a] at CEFR level {***level***} showing how you use the **{*pos* if verb/noun/adverb/adjective; else 'word' or 'expression'}** "**{*item*}**" ({***domain***}) :

a. The reason for 'sentence' to be used rather than 'example', even though some of the reference examples consist of more than a single sentence, is that using 'example' tends to result in the rendition of extensive explanations instead of or in addition to an example of use. This problem does not persist with the one-shot setting, for which therefore the word 'example' is used instead.

One-shot setting for Gemini :

> Please provide an example (between {***lower***}[a] and {***upper***} words at CEFR level {***level***} showing how you use the **{*pos* if verb/noun/adverb/adjective; else 'word' or 'expression'}** "**{*item*}**" ({***domain***}) :
>
> Example : {**reference_example**}

a. 'Lower' and 'upper' denote a range of example lengths, which differs for the different CEFR levels (8 to 43 words for B1 and 20 to 87 words for B2). The ranges are defined as +/- 1.5 standard deviations from the average value per level. This value as well as the addition of information about length itself was decided upon following a process of trial and error based on the behaviour of 20 sample examples in comparison to the reference's counterparts.

Appendix 3 : Features Used in Corpus Comparison

Length-Based	total number of examples in the sample total number of words in the sample average/min/max/SD number of words per sentence average/min/max/SD number of syllables per sentence average/min/max/SD number of letters per word average/min/max/SD number of syllables per word
Morphosyntactic	average/min/max/SD number of noun phrases per sentence average/min/max/SD number of non-stem words per s-ce percentage of sentences ending in question mark percentage of sentences ending in exclamation mark average/min/max/SD number of punctuation signs per s-ce morphological richness
Lexico-Semantic	average/min/max/SD number of verbs per sentence average/min/max/SD number of adj. and adv. per s-ce average/min/max/SD number of 1st-person pronouns per s-ce average/min/max/SD number of proper nouns per sentence percentage of words not present in the Dale-Chall list percentage of hapax legomena type-to-token ratio (word-based) type-to-token ratio (lemma-based) average concreteness 10 most frequent words (excluding stop words) 10 most frequent words (including stop words)
Discourse-Related	average/min/max/SD number of pronouns per sentence average/min/max/SD cosine distance between sentences average/min/max/SD % of anaphora-denoting words per sentence

Appendix 4 : Detailed Results

Entire Sample

Feature	Reference	Web-Crawled	Mistral	Vicuna	Gemini	Gemini : one-shot
Total # examples in sample	244	244	244	244	244	**244**
Total # words in sample	9823	7269	4615	4852	4160	**10366**
Avg. # words / s-ce	*13.59*	*14.93**[l]*	*9.34***	*9.6***	*8.51****	***12.26****
Min. # words / s-ce	3	7	4	8	8	**5**
Max. # words / s-ce	58	69	36	30	36	**44**
SD # words / s-ce	8.59	11.42	4.84	5.04	4.61	**5.77**
Avg. # syllables / s-ce	*21.52*	*24.47****	*15.33*	*15.93*	*14.65**	***20.47***
Min. # syllables / s-ce	2	12	4	10	14	**8**
Max. # syllables / s-ce	93	120	63	68	57	**84**
SD # syllables / s-ce	14.82	20.5	8.77	9.51	8.61	**10.97**
Avg. # letters / word	*5.29*	*5.38*	*5.3*	*5.43*	*5.6**	***5.53***
Min. # letters / word	1	1	1	1	1	**1**
Max. # letters / word	21	23	15	17	17	**20**
SD # letters / word	2.89	3.01	2.94	3.02	3.07	**3.01**
Avg. # syllables / word	*1.58*	*1.64**	*1.64****	*1.66****	*1.72****	***1.67****
Min. # syllables / word	0	0	1	1	1	**0**
Max. # syllables / word	7	8	6	6	6	**6**
SD # syllables / word	0.91	0.99	0.94	0.96	1	**0.96**

Avg. # noun phrases / s-ce	*5.87*	*8.16****	*5.56*	*5.53*	*4.92****	***5.39****
Min. # noun phrases / s-ce	0	2	1	2	2	**1**
Max. # noun phrases / s-ce	18	19	10	10	11	**14**
SD # noun phrases / s-ce	2.71	3.4	1.59	1.68	1.49	**1.9**
Avg. % non-stem words / s-ce	*33.56*	*31.56****	*35.14****	*35.59****	*36.36****	***36****
Min. % non-stem words / s-ce	0	0	11.76	11.11	9.1	**6.25**
Max. % non-stem words / s-ce	70	56.25	65	70	64.29	**83.33**
SD % non-stem words / s-ce	11.17	9.18	9.82	10.07	9.95	**11.15**
% s-ces ending in "?"	0.63	1.62	0	0	0	**0**
% s-ces ending in "!"	0.42	0	0.4	0	0	**0**
Avg. # punct. signs / s-ce	*1.56*	*2.7*	*0.77****	*0.99****	*0.75****	***1.25****
Min. # punct. signs / s-ce	0	0	0	0	0	**0**
Max. # punct. signs / s-ce	6	7	4	3	4	**6**
SD # punct. signs / s-ce	0.55	0.77	0.32	0.35	0.32	**0.48**
Morphological richness	0.02	0.02	0.02	0.02	0.02	**0.02**
Avg. # verbs / s-ce	*2.45*	*3.83*	*2.54****	*2.44****	*2.27****	***2.53***
Min. # verbs / s-ce	0	0	0	0	0	**0**
Max. # verbs / s-ce	8	11	7	7	5	**8**
SD # verbs / s-ce	1.53	1.84	1.08	1.26	1.06	**1.14**
Avg. # adj. and adv. / s-ce	*2.96*	*4.13****	*2.21****	*2.31****	*2.29****	***2.52****
Min. # adj. and adv. / s-ce	0	0	0	0	0	**0**

Max. # adj. and adv. / s-ce	10	14	10	7	8	**9**
SD # adj. and adv. / s-ce	1.97	2.62	1.48	1.46	1.41	**1.56**
Avg. # 1st-person pron. / s-ce	*0.1*	*0.12*	*0.39***	*0.11*	*0.06***	***0.07****
Min. # 1st-person pron. / s-ce	0	0	0	0	0	**0**
Max. # 1st-person pron. / s-ce	2	4	4	2	3	**4**
SD # 1st-person pron. / s-ce	0.38	0.41	0.75	0.38	0.3	**0.36**
Avg. # proper nouns / s-ce	*0.9*	*1.46*	*0.06****	*0.32****	*0.1****	***0.27*****
Min. # proper nouns / s-ce	0	0	0	0	0	**0**
Max. # proper nouns / s-ce	14	12	3	15	7	**10**
SD # proper nouns / s-ce	1.84	2.44	0.31	1.56	0.54	**0.94**
% words not in Dale-Chall list	45.21	45.38	42.25	45.22	48.34	**47.41**
% hapax legomena	16.13	22.56	16.25	18.18	19.75	**14.14**
Type-to-token ratio (words)	0.26	0.33	0.27	0.29	0.3	**0.24**
Type-to-token ratio (lemmas)	0.25	0.31	0.25	0.27	0.29	**0.23**
Average concreteness	2.46	2.36	2.44	2.44	2.42	**2.41**
10 most frequent words (excl. stop words)	water, climate, change, species, world, people, new, plants, global, could	would, could, said, people, water, also, international, may, must, new	crop, soil, crops, farmers, scientists, agriculture, water, yields, new, order	soil, crop, farmers, agriculture, crops, yields, practices, water, agricultural, climate	crop, soil, new, scientists, farmers, crops, practices, yields, scientist, sustainable	**water, soil, farmers, crops, crop, plant, species, food, practices, yields**
10 most frequent words (incl. stop words)	the, of, and, to, in, a, is, that, are, for	the, of, and, to, in, a, that, for, be, is	the, to, of, and, in, a, that, I, for, can	to, the, of, and, in, a, that, is, can, as the, of, to, a, and, in, for, crop, soil	the, of, to, a, and, in, for, crop, soil	**the, of, to, and, in, a, for, is, that, are**
Avg pron. / s-ce	*0.88*	*1.27*	*1.06***	*0.8***	*0.5****	***0.73****
Min. pron. / s-ce	0	0	0	0	0	**0**
Max. pron. / s-ce	5	6	5	6	4	**4**

SD pron. / s-ce	1.03	1.33	1.1	0.92	0.78	**0.93**
Avg. % anaphora words / s-ce	*20.49*	*10.78*	*10.47*	*10.43*	*13.42****	***24.59***
Min. % anaphora words / s-ce	0	0	0	0	0	**0**
Max. % anaphora words / s-ce	42.86	30	37.5	30.77	30.77	**31.25**
SD % anaphora words / s-ce	6.55	5.64	6.97	6.53	6.93	**6.8**
Avg. cos. distance btwn s-ces	*0.14*	*0.1****	*0.18****	*0.17****	*0.18****	***0.15****
Min. cos. distance btwn s-ces	-0.25	-0.25	-0.25	-0.32	-0.26	**-0.27**
Max. cos. distance btwn s-ces	0.72	0.78	0.85	0.9	0.87	**0.89**
SD cos. distance btwn s-ces	0.13	0.12	0.17	0.16	0.16	**0.15**

Comparison between the investigated corpora based on a sample of textual features. Features in *italics* have been tested for statistical significance, and the extent of the significance is marked with *, ** and *** from lowest to highest. The one-shot Gemini corpus is marked with **bold** to denote its highest global similarity to the reference corpus.

Per Level : B1 (domain "Science")

Feature	Reference	Web-Crawled	Mistral	Vicuna	Gemini	Combined LLM
Total # examples in sample	132	132	132	132	132	132
Total # words in sample	3402	3987	2358	2421	1976	2824
Avg. # words / s-ce	*11.01*	*15.1****	*8.8**	*9.03*	*7.46****	*9****
Min. # words / s-ce	3	7	9	8	8	5
Max. # words / s-ce	42	69	31	20	33	38
SD # words / s-ce	7.52	12.03	4.44	5.27	3.51	5.33
Avg. # syllables / s-ce	*17.81*	*24.62****	*14.66*	*14.93*	*12.72**	*14.71****
Min. # syllables / s-ce	2	12	10	11	14	8
Max. # syllables / s-ce	83	120	63	56	48	65
SD # syllables / s-ce	13.93	20.85	8.37	9.65	6.39	9.78
Avg. # letters / word	*3.56*	*5.35*	*5.27*	*5.35*	*5.45*	*5.35*
Min. # letters / word	1	1	1	1	1	1
Max. # letters / word	20	19	15	16	17	20
SD # letters / word	2.95	3	3.01	2.99	3.05	3
Avg. # syllables / word	*1.62*	*1.63*	*1.67*	*1.65*	*1.71**	*1.64**
Min. # syllables / word	0	0	1	1	1	0
Max. # syllables / word	7	8	6	6	6	6
SD # syllables / word	0.95	0.97	0.96	0.95	1	0.95

Avg. # noun phrases / s-ce	*5.51*	*8.01***	*5.26*	*5.3*	*4.38****	*5.14***
Min. # noun phrases / s-ce	0	2	2	2	2	1
Max. # noun phrases / s-ce	14	19	9	10	8	11
SD # noun phrases / s-ce	2.49	3.5	1.38	1.74	1.19	1.87
Avg. % non-stem words / s-ce	*34.3*	*31.01*	*34.37****	*34.41****	*34.72****	*34.1*
Min. % non-stem words / s-ce	0	0	11.76	11.11	9.09	8.33
Max. % non-stem words / s-ce	61.54	56.25	61.54	70	57.14	55
SD % non-stem words / s-ce	10.61	9.63	10.1	10.31	9.35	9.95
% s-ces ending in "?"	0	2.22	0	0	0	0
% s-ces ending in "!"	1.14	0	0	0	0	0
Avg. # punct. signs / s-ce	*1.33*	*2.71****	*0.63****	*0.88***	*0.46****	*0.91****
Min. # punct. signs / s-ce	0	0	0	0	0	0
Max. # punct. signs / s-ce	6	7	3	3	3	6
SD # punct. signs / s-ce	0.54	0.79	0.28	0.33	0.24	0.4
Morphological richness	0.02	0.02	0.02	0.02	0.02	0.02
Avg. # verbs / s-ce	*2.17*	*3.84****	*2.41*	*2.35*	*2.03****	*2.21*
Min. # verbs / s-ce	0	0	0	0	0	0
Max. # verbs / s-ce	7	11	7	7	4	7
SD # verbs / s-ce	1.39	1.95	1.03	1.25	0.86	1.23
Avg. # adj. and adv. / s-ce	*2.78*	*4.09*	*2.04****	*2.13****	*1.97****	*2.07****
Min. # adj. and adv. / s-ce	0	0	0	0	0	0

Max. # adj. and adv. / s-ce	9	14	10	7	6	6
SD # adj. and adv. / s-ce	1.81	2.67	1.45	1.43	1.22	1.32
Avg. # 1st-person pron. / s-ce	*0.12*	*0.16*	*0.55****	*0.14*	*0.1*	*0.08*
Min. # 1st-person pron. / s-ce	0	0	0	0	0	0
Max. # 1st-person pron. / s-ce	2	4	4	2	3	2
SD # 1st-person pron. / s-ce	0.42	0.51	0.88	0.41	0.39	0.38
Avg. # proper nouns / s-ce	*0.71*	*1.64**	*0.1****	*0.36****	*0.11****	*0.2****
Min. # proper nouns / s-ce	0	0	0	0	0	0
Max. # proper nouns / s-ce	7	12	3	15	2	5
SD # proper nouns / s-ce	1.34	2.72	0.38	1.45	0.36	0.69
% words not in Dale-Chall list	46.06	45.78	41.65	44.4	46.51	45.56
% hapax legomena	24.06	29.82	23.3	26.57	28.11	26.84
Type-to-token ratio (words)	0.36	0.4	0.35	0.38	0.39	0.38
Type-to-token ratio (lemmas)	0.34	0.38	0.33	0.35	0.37	0.36
Average concreteness	2.47	2.39	2.38	2.41	2.37	2.41
10 most frequent words (excl. stop words)	climate, change, water, people, areas , earthquake, greenhouse, earth, world, atmosphere	would, could, time, may, said, water, help, new, using, must	scientists, science, study, new, experiment, behavior, scientific, research, climate, understand	scientists, scientific, research, study, field, climate, science, discovery, certain, understand	scientists, new, scientist, experiment, earth, research, data, chemical, scientific, study	water, experiment, scientists, chemical, new, used, species, different, earch, causing
10 most frequent words (incl. stop words)	the, of, and, in, to, a, is, are, as, can	the, of, to, and, in, a, that, for, be, is	the, of, to, in, a, and, I, that, scientists, is	to, the, and, in, of, a, is, as, that, it	the, of, to, a, scientists, in, and, new, for, that	the, of, to, in, a, and, is, that, are
Avg pron. / s-ce	*0.67*	*1.27***	*1.25****	*0.77*	*0.47**	*0.65**
Min. pron. / s-ce	0	0	0	0	0	0
Max. pron. / s-ce	5	6	5	4	3	4

SD pron. / s-ce	0.99	1.39	1.11	0.91	0.69	0.91
Avg. % anaphora words / s-ce	*12.95*	*10.63*	*11.7**	*11.41*	*15.9****	*16.53**
Min. % anaphora words / s-ce	0	0	0	0	0	0
Max. % anaphora words / s-ce	27.27	30	37.5	30.77	30.77	31.25
SD % anaphora words / s-ce	6.41	5.73	6.82	6.76	7.05	6.59
Avg. cos. distance btwn s-ces	*0.13*	*0.09****	*0.14****	*0.13*	*0.16****	*0.11****
Min. cos. distance btwn s-ces	-0.22	-0.25	-0.25	-0.25	-0.26	-0.27
Max. cos. distance btwn s-ces	0.71	0.64	0.84	0.8	0.87	0.78
SD cos. distance btwn s-ces	0.14	0.11	0.14	0.13	0.14	0.11

Per Level : B2 (domain "Agronomy")

Feature	Reference	Web-Crawled	Mistral	Vicuna	Gemini	Gemini : one-shot
Total # examples in sample	112	112	112	112	112	112
Total # words in sample	6421	3282	2257	2431	2184	7542
Avg. # words / s-ce	*15.51*	*14.72****	*9.99*	*10.26*	*9.75*	*14.21****
Min. # words / s-ce	3	11	4	9	11	7
Max. # words / s-ce	58	64	36	30	36	44
SD # words / s-ce	9.09	10.69	5.02	4.64	4.57	5.85
Avg. # syllables / s-ce	*24.29*	*24.29****	*16.12*	*17.06*	*16.94*	*23.91*
Min. # syllables / s-ce	3	16	4	10	19	9
Max. # syllables / s-ce	93	106	57	68	57	84
SD # syllables / s-ce	15.28	20.16	9.02	9.15	8.72	11.22
Avg. # letters / word	*5.26*	*5.4*	*5.33*	*5.5**	*5.75****	*5.6****
Min. # letters / word	1	1	1	1	1	1
Max. # letters / word	21	23	15	17	17	20
SD # letters / word	2.85	3.01	2.86	3.05	3.09	3.03
Avg. # syllables / word	*1.57*	*1.65***	*1.61**	*1.66****	*1.74****	*1.68****
Min. # syllables / word	0	0	1	1	1	0
Max. # syllables / word	7	8	5	5	6	6
SD # syllables / word	0.89	1.02	0.91	0.97	1	0.97

Avg. # noun phrases / s-ce	*6.09*	*8.34****	*5.91*	*5.78*	*5.57*	*5.49*
Min. # noun phrases / s-ce	0	3	1	2	2	1
Max. # noun phrases / s-ce	18	17	10	9	11	14
SD # noun phrases / s-ce	2.81	3.28	1.74	1.58	1.55	1.91
Avg. % non-stem words / s-ce	*31.64*	*32.22****	*35.95****	*36.77****	*37.83****	*36.71****
Min. % non-stem words / s-ce	0	14.71	12.5	13.04	15.79	6.25
Max. % non-stem words / s-ce	70	56.25	65	61.11	64.29	83.33
SD % non-stem words / s-ce	11.41	8.61	9.45	9.71	10.41	11.51
% s-ces ending in “?”	1	0.89	0	0	0	0
% s-ces ending in “!”	0	0	0.87	0	0	0
Avg. # punct. signs / s-ce	*1.7*	*2.69****	*0.93****	*1.1****	*1.1****	*1.4*
Min. # punct. signs / s-ce	0	0	0	0	0	0
Max. # punct. signs / s-ce	6	5	4	3	4	6
SD # punct. signs / s-ce	0.56	0.76	0.36	0.38	0.39	0.51
Morphological rich-ness	0.02	0.02	0.02	0.02	0.02	0.02
Avg. # verbs / s-ce	*2.61*	*3.81****	*2.7****	*2.54****	*2.56****	*2.67**
Min. # verbs / s-ce	0	0	0	0	0	0
Max. # verbs / s-ce	8	10	6	6	5	8
SD # verbs / s-ce	1.58	1.69	1.13	1.27	1.2	1.5
Avg. # adj. and adv. / s-ce	*3.07*	*4.17****	*2.42****	*2.5****	*2.67****	*2.71*
Min. # adj. and adv. / s-ce	0	0	0	0	0	0

Max. # adj. and adv. / s-ce	10	14	8	6	8	9
SD # adj. and adv. / s-ce	2.04	2.58	1.5	1.48	1.53	1.62
Avg. # 1st-person pron. / s-ce	*0.09*	*0.06**	*0.2*	*0.08**	*0.02****	*0.06**
Min. # 1st-person pron. / s-ce	0	0	0	0	0	0
Max. # 1st-person pron. / s-ce	2	1	2	2	1	4
SD # 1st-person pron. / s-ce	0.35	0.24	0.5	0.35	0.13	0.35
Avg. # proper nouns / s-ce	*1.05*	*1.24***	*0.02****	*0.28****	*0.11****	*0.3****
Min. # proper nouns / s-ce	0	0	0	0	0	0
Max. # proper nouns / s-ce	14	10	2	15	7	10
SD # proper nouns / s-ce	2.07	2.04	0.19	1.67	0.7	1.03
% words not in Dale-Chall list	44.76	44.91	42.89	46.03	50	48.11
% hapax legomena	20.15	27.51	18.91	21.35	21.89	15.77
Type-to-token ratio (words)	0.31	0.39	0.3	0.32	0.33	0.27
Type-to-token ratio (lemmas)	0.29	0.37	0.29	0.31	0.31	0.25
Average concreteness	2.45	2.36	2.51	2.48	2.47	2.41
10 most common words (excl. stop words)	water, species, climate, change, could, plants, world, many, help, new	could, people, said, would, also, international, humanitarian, countries, must, called	crop, soil, crops, farmers, agriculture, yields, water, farmer, yield, conditions	crop, soil, farmers, agriculture, crops, yields, practices, agricultural, sustainable, water	crop, soil, farmers, crops, practices, yields, sustainable, agricultural, agronomist, farmer	water, soil, farmers, crop, crops, yields, plant, farming, practices, food
10 most frequent words (incl. stop words)	the, of, to, and, in, a, is, that, are, for	the, and, of, to, in, that, a, for, it, with	the, to, and, of, in, crop, soil, crops, a, can	to, the, and, in, of, crop, soil, can, that, are	the, to, of, and, in, a, for, crop, soil, farmers	the, end, to, of, a, in, for, that, their, can
Avg pron. / s-ce	*1*	*1.27****	*0.84****	*0.84****	*0.54****	*0.77*
Min. pron. / s-ce	0	0	0	0	0	0
Max. pron. / s-ce	4	5	5	5	4	4

SD pron. / s-ce	1.03	1.27	1.06	0.93	0.88	0.94
Avg. % anaphora words / s-ce	*10.5*	*10.97*	*9.01****	*9.27****	10.51	*10.75*
Min. % anaphora words / s-ce	0	0	0	0	0	0
Max. % anaphora words / s-ce	42.86	27.27	30.77	27.27	25	30
SD % anaphora words / s-ce	6.6	5.53	6.91	5.96	5.53	6.58
Avg. cos. distance btwn s-ces	*0.16*	*0.12****	*0.38****	*0.32****	*0.35****	*0.24****
Min. cos. distance btwn s-ces	-0.21	-0.22	-0.05	-0.22	-0.06	-0.17
Max. cos. distance btwn s-ces	0.72	0.78	0.85	0.89	0.87	0.89
SD cos. distance btwn s-ces	0.13	0.13	0.13	0.17	0.14	0.16