Generating Contexts for ESP Vocabulary Exercises with LLMs

Iglika Nikolova-Stoupak[1*], **Serge Bibauw**[2], **Amandine Dumont**[3], **Françoise Stas**[3], **Patrick Watrin**[1], **Thomas François**[1]

[1] CENTAL, Université catholique de Louvain, Belgium,
[2] GIRSEF, Université catholique de Louvain, Belgium
[3] ILV, Université catholique de Louvain, Belgium
[*] iglika.nikolova@uclouvain.be

Abstract

The current paper addresses the need for language students and teachers to have access to a large number of pedagogically sound contexts for vocabulary acquisition and testing. We investigate the automatic derivation of contexts for a vocabulary list of English for Specific Purposes (ESP). The contexts are generated by contemporary Large Language Models (namely, Mistral-7B-Instruct and Gemini 1.0 Pro) in zero-shot and few-shot settings, or retrieved from a web-crawled repository of domain-relevant websites. The resulting contexts are compared to a professionally crafted reference corpus based on their textual characteristics (length, morphosyntactic, lexico-semantic, and discourse-related). In addition, we annotated the automatically derived contexts regarding their direct applicability, comprehensibility, and domain relevance. The 'Gemini, zero-shot' contexts are rated most highly by human annotators in terms of pedagogical usability, while the 'Mistral, few-shot' contexts are globally closest to the reference based on textual characteristics.

1 Introduction

The development of a wide vocabulary is a fundamental component of foreign language acquisition as it underpins the development of all other language skills (Ardasheva et al., 2019; Gorjian et al., 2011). To pursue this aim, learners are typically encouraged to exploit multiple strategies, such as studying from traditional mono- or bilingual vocabulary lists or making use of technology-based resources such as digital flashcards or vocabulary learning apps (Restrepo Ramos, 2015).

Research shows that new vocabulary items are better acquired when encountered in authentic and informative contexts (Huckin and Coady, 1999; Restrepo Ramos, 2015; Godwin-Jones, 2018). However, looking for or coming up with high-quality contexts, especially more advanced and specialised ones, presents a serious challenge to teaching professionals in terms of time and effort. Therefore, the use of contemporary Natural Language Processing (NLP) techniques to come up with a large number of pedagogically sound contexts would present a significant benefit to both teachers and learners.

Against this backdrop, this paper presents our detailed experiments in deploying NLP methods to generate or retrieve contexts to help the acquisition of specialised English vocabulary by French-speaking university students reading science and agronomy. We used two Large Language Models (LLMs) of different sizes, namely Mistral-7B-Instruct and Gemini 1.0 Pro (in the context of both a zero-shot and a few-shot setting) and a custom-made web-based scientific corpus to produce context sentences for a predefined vocabulary list of 100 items belonging to CEFR levels B1-B2 in those two specialised domains. Our ultimate goal is to use the issuing contexts in the creation of exercises of the 'gapfill' and 'multiple-choice' types (see Fig. 1).

In this context, this paper addresses the three following research questions:

1. Which derivation method (web retrieval or LLM-generated) results in contexts for ESP vocabulary learning that are closer to professionally crafted ones in terms of textual char-

Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin and Thomas François. Generating Contexts for ESP Vocabulary Exercises with LLMs. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*. Linköping Electronic Conference Proceedings 211: 153–175.

1. Climate models have traditionally shown considerable inaccuracy in their simulations of the Arctic. This **sh**_____ is particularly troubling nowadays, because the Arctic is the region expected to undergo the most extreme climate changes in the future.

2. Climate models have traditionally shown considerable inaccuracy in their simulations of the Arctic. This _____ is particularly troubling nowadays, because the Arctic is the region expected to undergo the most extreme climate changes in the future.

a. cluster **b. shortcoming** c. assertion d. insight e. endeavour

Figure 1: Examples of relevant 'gapfill' (1) and 'multiple-choice' (2) questions.

acteristics?

2. To what extent is it possible to guarantee the pedagogical quality of the issued contexts and their ready application in the classroom?

3. Is there a perceivable correlation between the contexts' textual characteristics and their pedagogical qualities as evaluated by teaching professionals?

The paper is organised as follows: Section 2 discusses related work regarding the pedagogical qualities of educational texts, automatic derivation of teaching materials, as well as their evaluation, with a particular focus on materials for the acquisition of EFL vocabulary. Section 3 explains our methodology for assembling and evaluating the examined corpora, and Section 4 presents the results of our experiments. We discuss our main findings in Section 5 and finally offer a conclusion and future directions in Section 6.

2 Background

2.1 Pedagogical Characteristics of Texts

There exists a variety of theories and perspectives when it comes to the definition of what makes a text suitable for a pedagogical setting, particularly in the context of foreign language learning. Siregar and Purbani (2024) draw attention to a number of narrow grammatical features as a guarantee for pedagogical suitability, such as the lack of nominalisations and extensive modifiers and the use of simpler patterns, such as *noun + preposition* or single clauses. Pedagogical qualities may also be dependent on the specific classroom addressed. Targeting younger learners, Morais and Neves (2010) underline the importance of interdisciplinarity in learning materials and tasks. Yet, most researchers agree that the essential prerequisite for any input in language acquisition is that it should be "contextualised and comprehensible" (Tomlinson, 2012, 156).

Much emphasis has been placed on a text's *authenticity* as a pedagogical quality. A text is seen as authentic if it has been produced to serve a social purpose rather than a pedagogical one (Little et al., 1989). As a document's feature, authenticity has, hence, commonly been equated to a lack of adaptation, to the retaining of a text's original goal or context (Besse, 1981; Crossley et al., 2007). Yet, the superiority of authentic texts is still a subject of debate. Text simplification, for instance, has been shown to provide clear pedagogical advantages in textual characteristics (Crossley et al., 2007) and in comprehension and vocabulary learning effects (Rets and Rogaten, 2021). The emergence of generative AI also opens new debates on what qualifies as authentic, as such applications produce texts that are neither pedagogical nor the product of genuine human communication.

2.2 Automatic Derivation of Teaching Materials

The large amount of available data and the automatisation opportunities that recent technology offers have been used extensively in the composition and presentation of teaching materials, particularly in the English as a Foreign Language (EFL) classroom. Various types of texts are derived from the web and typically adapted for use in a specific learning setting (Litman, 2016; Meurers et al., 2010). For instance, Heilman et al. (2008) gather a web-based textual corpus meant for vocabulary and reading practice as well as devise a user-friendly system (REAP Search) that enables the selection of elements from the corpus based on a list of relevant constraints.

In the past few years, LLMs have also been exploited in the language classroom due to their revolutionary ability to produce language based on personalised instructions. Expectedly, due to

its popularity and ease of access, ChatGPT has been receiving particular attention. A number of experimental studies have been conducted internationally in an attempt to define and estimate the chatbot's potential to aid students in the ESL classroom. Following interaction with ChatGPT, learners of various age and proficiency levels are commonly discovered to have been motivated by the tool; furthermore, their academic results have been objectively improved, notably in the field of vocabulary acquisition, thanks to activities such as conversational practice and work with automatically generated text (Young and Shishido, 2023a,b; Shaikh et al., 2023; Songsiengchai et al., 2023; Aktay and Uzunoglu, 2023; Lou, 2023).

2.3 Evaluation of Automatically Derived Teaching Materials

Jeon and Lee (2023) sum up LLMs' applicability to language education as belonging to four discrete roles, namely interlocutor, content provider, teaching assistant, and evaluator. As per their last role, LLMs are claimed to be able to automatically evaluate the quality of student- and teacher-produced materials, as well as of automatically generated ones. Yet, such an evaluation by LLMs has not been substantially addressed due to its qualitative nature, and consequently, more traditional NLP techniques, especially related to readability or, otherwise, textual complexity in its different aspects, are typically applied to estimate textual quality and/or suitability. For instance, Loiseau et al. (2005) proposed an NLP-based system for pedagogical indexation where, upon insertion of a text or extract and indication of the intended learners' level, its difficulty is estimated, and elements that may need to be adapted, such as complex grammatical tenses or vocabulary items, are highlighted. Aiming at consistent and large-scale evaluation of adapted internet materials, Hussin et al. (2010) performed a correlation analysis between the difficulty of texts as estimated by teachers and their readability characteristics, discovering statistical significance in relation to average sentence length, average word length and the coverage of the first 2000 high-frequency words.

Relevant human-based counterparts of generated materials have also been utilised as ground truth against which to evaluate them. For instance, Yunjiu et al. (2022) specifically addressed the evaluation of vocabulary exercises; more specifically, in Chinese as a target language. They evaluated the quality of AI-generated distractors (non-correct answers) for multiple-choice questions based on a combination of semantic and visual similarity to the correct answer. Results and qualitative reflections of the test takers suggested that the automatically generated distractors are more complicated, possibly for reasons including the semantic similarity between them and their absence from textbooks used by the students. In a study related to the present one, Nikolova-Stoupak et al. (2024) generated and retrieved a number of contexts around ESP vocabulary list items and evaluated them based on their closeness to a gold standard of professionally crafted contexts in terms of a number of atomic readability-related features. Generated teaching materials for vocabulary acquisition have also been evaluated quantitatively in terms of compactness or informativeness. Paddags et al. (2024) generated sentences aimed at the teaching of Danish vocabulary using a few-shot LLM setting and consequently evaluated their quality based on their density in terms of the number of target words (based on a defined vocabulary list) that fit into a single sentence.

3 Methods

We conducted a series of experiments in deploying NLP methods to generate and retrieve contexts around a predefined vocabulary list of 100 items belonging to CEFR levels B1-B2 and the domains of general science and agronomy[1]. Each item is associated with a gold standard context as handpicked by teaching professionals and previously used in a classroom for testing purposes (gapfill or multiple-choice questions). In particular, we generated contexts using two Large Language Models (LLMs) of different sizes, Mistral-7B-Instruct and Gemini 1.0 Pro, in both a zero-shot and a few-shot setting. In addition, we composed a corpus of scientific articles from relevant web sources and formulated a pipeline to extract relevant context sentences from them.

An important part of our work was to devise methods that guarantee that the derived contexts are of high educational quality and are thus directly applicable in an ESP classroom setting. Via

[1]the items were selected based on a larger pre-selection verified by teaching professionals; a balance between parts of speech was sought

hand-crafted rules, we ensured that the derived contexts resemble the gold standard defined by Nikolova-Stoupak et al. (2024) in terms of linguistic characteristics (as represented by common readability features). Additionally, we limited output to the appropriate scientific domain and CEFR level with the help of prompt engineering and classifier-based filters. The contexts issued from the different derivation methods were then manually annotated by experienced teachers of ESP from the Catholic University of Louvain in terms of their educational quality. Using insights from this human evaluation, we classified the contexts and, by extension, the methods behind their derivation, discussing their qualities and drawbacks and drawing conclusions about the interdependence between their automatable linguistic characteristics and their pedagogical qualities.

This section elaborates on the automatic derivation of the corpora (for an illustration of the process, see Figure 2) as well as on the methods applied in their evaluation.

3.1 Retrieval of Web-Crawled Contexts

Firstly, all accessible articles from a list of thematic websites as defined by a team of ESP teachers (see Appendix 1: List of Crawled Websites) were retrieved through web-crawling Python tools, such as *beautifulsoup4*[2] and *newspaper*[3] and shaped into a database along with metadata including the textual format, date, the source webpage and its associated domain[4]. The derived text underwent a simple cleaning pipeline, such as the removal of non-alphanumeric symbols and non-English text. Context sentences associated with the predefined vocabulary list were then extracted from the database using a pipeline of hand-crafted rules. The articles were surveyed to determine the occurrence of the target vocabulary items or their alternative forms. When the search form was mapped, consistency was sought with the item's domain and part of speech.

Several filters were then applied, ensuring that the target item is present in the sentence only a single time, that the sentence can be considered as scientific, that its CEFR level closely matches the intended level, and, eventually, that its linguistic characteristics[5] resemble those of a set of professionally crafted contexts sampled from the reference dataset of Nikolova-Stoupak et al. (2024). More precisely, this proximity was measured as Euclidean distance over the set of features and data points of up to two standard deviations from the values computed on the reference corpus were retained[6].

3.1.1 CEFR Level Classifier

It was important to guarantee that the contexts that were retrieved closely matched the CEFR level associated with the vocabulary list items that they were mapped with. Upon experimentation with established and readily available tools designed for estimation of the CEFR level of texts, such as *Textinspector*[7] and *English CEFR Level Predictor*[8], it was observed that these solutions do not work well when faced with text that is a single sentence of length. Therefore, a custom classifier was trained to determine the extracted sentences' CEFR levels. We built a corpus of sentences annotated with their CEFR level by concatenating Arase et al. (2022)'s WikiAuto- and SCoRE-based corpora, which were annotated by two experienced teaching experts[9] and the sentences available through the *English Profile* website (Salamoura and Saville, 2010), which are originally taken from the Cambridge Learner Corpus[10] and exemplify discrete CEFR levels with their characteristics. We then used this corpus, which totalled 13,378 sentences, to finetune a BERT model

[2]Version 4.12.3; https://pypi.org/project/beautifulsoup4/

[3]Version 0.2.8; https://pypi.org/project/newspaper3k/

[4]Among the three following domains: science, agronomy, and technology.

[5]The set of characteristics that we considered in this work is: the number of words, the number of letters per word, the number of punctuation signs, the number of noun phrases, the percentage of non-stem words, the number of first-person pronouns, the number of proper nouns, the number of pronouns, and the number of anaphora-denoting words. Please refer to Appendix 3: Features Used in Corpus Comparison for details about these characteristics.

[6]For some features, the value was increased or reduced based on observations. Sentence length was thus limited to 1.5 standard deviations from the reference, and the percentage of non-stem words was relaxed to 3 standard deviations. When present, negative values were rounded to 0.

[7]https://textinspector.com/api-developers/

[8]https://github.com/AMontgomerie/CEFR-English-Level-Predictor

[9]where the two annotators' estimations differed, we took the higher CEFR level as it is less problematic for students to be provided with text that is slightly below their current level

[10]https://www.sketchengine.eu/cambridge-learner-corpus/

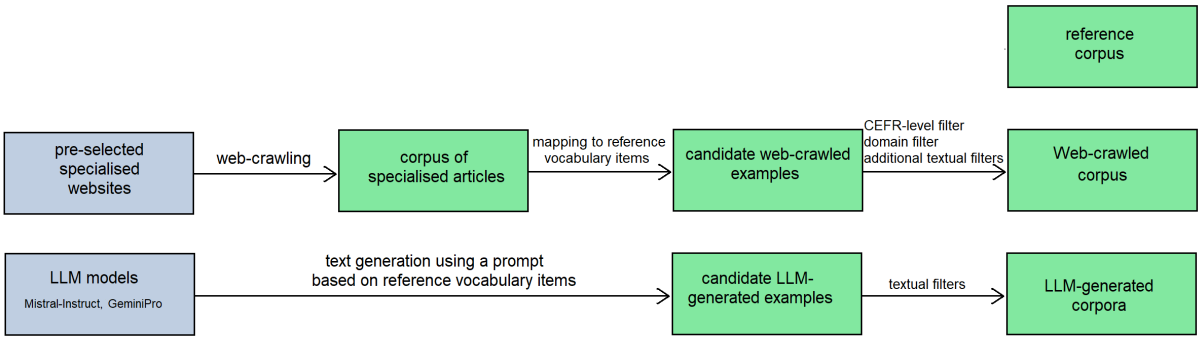

Figure 2: Collection procedure for the examined corpora

with a classification layer[11]. The derived classifier achieved 63% of accuracy[12], the majority of mistakes being associated with the marginal A1 and C2 proficiency levels, which are absent from our reference vocabulary list. When adjacent levels were considered, the accuracy went up to 98%. Given the qualitative nature of CEFR levels and the lack of full agreement between the used corpus annotators, candidate web-crawled sentences were retained if they belonged to the associated course's level or differed from it by a single level. Ultimately, only a small portion (around 10%) of the candidate sentences were discarded based on the CEFR-level filter.

3.1.2 Scientific Domain Classifier

As the web-crawled articles do not consist of scientific text in their entirety (e.g. there may be isolated informal sentences or even metadata within them), a binary SVM classifier model[13] was trained to label sentences that belong to a broad scientific domain. At first, the training corpus was composed of 2k scientific and 14k non-scientific sentences (2k 'law', 2k 'business', 2k 'sports', 2k 'world news', 2k 'law', 2k 'informal communication', and 2k 'literature'), taken at random from the following sources: respectively, PubMed[14] (the 'scientific' label); the Caselaw Access Project[15]; the AG News Classification Dataset's Business News, Sports News and World News subcorpora[16], Reddit's API[17], and an assembled corpus of full and abridged classical literary texts as freely available online. The classifier's performance was then tested on a random 100-sentence sample extracted from our web-crawled corpus, and a bias toward complex sentences, as well as an underrepresentation of certain scientific fields, such as chemistry, were detected. In order to improve the classifier, 1000 sentences with a length of up to 2 standard deviations from the reference value for the feature (as defined by Nikolova-Stoupak et al. (2024)), which were also manually confirmed to be scientific, were added to the training corpus's 'scientific' label. The newly derived classifier achieved 93% accuracy, and its performance was verified against the 100 manually labelled sentences and judged to act satisfactorily as a filter. The resulting classifier was used in the extraction of web-crawled context sentences, and non-scientific sentences (which turned out to be about one-third of the candidate sentences) were disregarded.

3.2 Generation of Contexts by LLMs

Two discrete contemporary LLMs were used for context generation: Mistral-7B-Instruct and Gemini 1.0 Pro. The former is a compact model whose performance compares to and occasionally surpasses that of LLaMA (Jiang et al., 2023), whilst the latter is a 600B variety of Gemini, a model that achieves state-of-the-art results in a number of key NLP tasks (Team et al., 2024) and is characterised with fast performance. For this experiment, Mistral was used through the 'LM studio' interface, and Gemini was accessed through the Google AI Studio developer tool, as freely available within a

[11] BERT was opted for in this task due to its strong language understanding and generation abilities

[12] Arase et al. (2022)'s associated classifier reaches a macro-F1 score of 84.5% as a result of elaborate techniques especially aimed at the correct recognition of sentences belonging to the rarer and marginal CEFR levels

[13] the choice of model was based on experiments with a few models that are strong in binary classification tasks

[14] https://pubmed.ncbi.nlm.nih.gov/

[15] https://case.law/docs/

[16] https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset

[17] https://www.reddit.com/wiki/api/

given quota at the time of writing. Following experiments, Mistral's temperature setting was adjusted to 0.8, as below this value, the output was highly homogeneous and commonly consisted of definitions of the target vocabulary. The experimental setup featured an 11th Gen Intel Core i7 CPU with 8 cores, and TigerLake-LP GT2 integrated GPU.

Both models were instructed to provide context examples based on the vocabulary list's items, parts of speech and domains in both a zero-shot setting and a few-shot setting. Within the latter, five examples of paired vocabulary items of various parts of speech and corresponding reference contexts as provided by teaching professionals were added to the prompts. For the full prompts utilised, please refer to Appendix 2: Prompts used for LLM Generation.

In addition, generation for a vocabulary item was iterated through until a number of conditions pertaining to the output were satisfied. As with the retrieval of web-crawled contexts, it was ensured that the target item was only present in the example a single time and that the example was of proximity to the gold standard sampled from Nikolova-Stoupak et al. (2024) as measured with Euclidean distance based on a selection of readability features[18]. We confirmed that the output was in English as well as the compatibility of the its part of speech and the absence of metatextual information (e.g. explanations of use) in addition to context examples.

3.3 Human Annotation

For the purpose of annotation, each of the five methods described above (generation with Mistral and Gemini in a zero-shot and few-shot setting and retrieval from a web-crawled database) was used to generate one context for each of the 100 words in our vocabulary list. This amounted to a total of 500 contexts, which were assigned unique IDs before being shuffled and information about their generation method being removed. Two ESP teachers with substantial experience were asked to evaluate the contexts for the following three pedagogical features: 'ready to use', 'comprehensible' and 'in-domain'. A 'ready to use' context was defined as being directly applicable for classroom use and assessment purposes without editing; 'comprehensibility' referred to a context being self-explanatory and understandable if encountered in isolation; finally, 'in-domain' meant that the contexts' field of specialisation is appropriate for students in the intended specialisation (i.e. science or agronomy). A Likert scale from 1 to 5 was utilised for the annotation, 5 signifying maximal possession of the quality in question. The annotators were also invited to leave comments in free text in relation to each of the evaluated contexts.

Initially, the annotators were given the first 100 contexts to annotate independently of each other, following which the inter-rater agreement between them was calculated. The 'ready to use' and 'comprehensible' categories are marked with moderate agreement according to Cohen's Quadratic Kappa but demonstrate good scores for exact agreement (respectively 65% and 87%). The 'in-domain' category comes with a low Cohen's Kappa value in combination with 97% exact agreement, a phenomenon caused by the heavily skewed ratings towards a maximal number of points for the category (Pontius Jr and Millones, 2011). As a next step, the annotators gathered to adjust the annotation guidelines and agree on a gold value for the items where their initial annotation differed. The remaining 400 contexts were then split between the two teachers to annotate.

3.4 Context Evaluation

The 500 contexts, as well as the 100 reference ones, were evaluated based on readability-related textual characteristics (see Appendix 3: Features Used in Corpus Comparison). An analysis identical to the one defined by Nikolova-Stoupak et al. (2024) followed. That is to say, firstly, the non-parametric Mann-Whitney U test was used to measure the significance of the difference between the reference corpus and contexts for each of the five collection methods (retrieval from a web-crawled corpus and generation by Mistral and Gemini in a zero-shot and few-shot setting). In turn, statistical significance was assigned to one of three levels, corresponding to p-values of 0.001, 0.01, and 0.05. In addition, the global distance between the reference corpus and each of the five context corpora was determined through the use of Euclidean distance between the totality of characteristics as having undergone min-max normalisation. The five associated derivation methods were ranked based on their closeness to the reference corpus. As

[18]The same ones as referred to in section 3.1.

an additional experiment, the examined vocabulary items were divided into CEFR levels B1 (56 items) and B2 (44 items) and all textual characteristics were evaluated once again in an attempt to reveal the derivation methods' sensitivity to the CEFR level at hand.

The derived corpora were also ranked based on their pedagogical qualities, as estimated by the human evaluation. For this purpose, each corpus was given a percentage value representing the number of points received for all evaluated categories assembled ('ready to use, 'comprehensible,' and 'in-domain') compared with the total number of points possible.

Ultimately, the two rankings were compared in an attempt to reveal a potential link between the contexts' linguistic and pedagogical qualities. It was assumed that the most highly rated method in the annotation process objectively has the highest pedagogical value.

4 Results

4.1 Automatic Evaluation

The corpus discovered to be globally closest to the reference one in terms of Euclidean distance based on all examined numeric textual characteristics is 'Mistral, few-shot' (3.82), followed by 'Gemini, few-shot' (4.86), 'Mistral, zero-shot' (4.99), 'Web-crawled' (5.46) and 'Gemini, zero-shot' (5.65). The 'Mistral, few-shot' model remains closest when the four categories of textual characteristics are considered separately, and the rest of the models mostly keep their place. The 'Mistral, zero-shot' model varies from second (for lexico-semantic and discourse-based characteristics) to fourth place (for length-based characteristics). The 'Web-crawled' corpus is closest to the reference in relation to length-based characteristics.

Table 1 shows a summary of the most relevant results of the corpus comparison based on atomic textual characteristics. For a comparison of all features, please refer to Appendix 4: Detailed Results of the Comparison between Corpora based on Textual Features.

The reference corpus is generally associated with the highest ranges (i.e. distances between the maximal and minimal values) as well as the highest standard deviation for continuous characteristics. The 'Mistral, few-shot' corpus often comes closest to the reference in these aspects (e.g. in relation to the number of words per sentence, the number of noun phrases per sentence, and the percentage of non-stem words per sentence).

The total number of words in the 'Mistral, few-shot' sample is closest to the reference and the only one larger. When length-based textual characteristics[19] as well as morphosyntactic characteristics[20] are considered, the 'Gemini, few-shot' corpus presents the least deviation from the reference. In the latter category, the 'Mistral, few-shot' corpus often comes closest to the reference, such as in terms of number of punctuation signs per sentence and the variety in end-of-sentence punctuation. The least statistical deviation when it comes to lexico-semantic characteristics is associated with the 'Mistral, zero-shot' corpus[21]. The most frequent words encountered in the 'Web-crawled' corpus strike as very generic and unrelated to the scientific domain compared to those in other corpora (e.g. 'would,' 'could,' 'said'). Within the 'Mistral, zero-shot' corpus, the personal pronoun 'I' is uniquely featured among the most frequent words when stop words are retained. Finally, discourse-related characteristics demonstrate little deviation from the reference, with the exception of those related to cosine distance, where statistical significance is smallest with the 'Web-crawled' sample. When subcorpora associated with CEFR level B1 are considered, the 'Mistral, few-shot' corpus demonstrates the lowest deviation from the reference corpus (only 4 features exhibiting statistical significance). In contrast, statistical significance is present in a minimum of 7 features for the others[22]. The two CEFR levels are also associated with different domains ('science' for B1 and 'agronomy' for B2), and this additional focus is reflected in the most used words for some of the corpora (e.g. the word 'scientists' is present for all LLM-based B1 subcorpora and the words 'crop' and 'soil' for the 'Gemini, zero-shot' and both Mistral B2 subcorpora).

4.2 Human Annotation

For a distribution of the values given to the corpora in the annotation in relation to the three character-

[19] The only (highly) significant deviation is for the average number of words per sentence.

[20] one significant deviation with moderate significance: the number of punctuation signs per sentence

[21] one instance of statistical significance of high value, for the number of proper nouns per sentence

[22] a number shared by the 'Web-crawled', 'Mistral, zero-shot' and 'Gemini, few-shot corpora'

Feature	Ref.	Web-crawled	Mistral, 0-shot	Mistral, f-shot	Gemini, 0-shot	Gemini, f-shot
words in sample	3787	2267	2823	**4091**	2345	2638
words / sentence	*13.33*	*11.11****	*11.67****	***13.73***	*11.17****	*11.32****
letters / word	*5.2*	*5.37*	*5.57****	***5.4****	*5.84****	*5.21*
noun phrases / sentence	*5.76*	*6.34**	*6.08*	***6.29****	*6.26**	*5.68*
non-stem words / s-ce	*31.91*	*34.2*	*38.06****	***35.2****	*40.09****	*33.28*
punctuation signs / s-ce	*1.51*	*0.98**	*1.06***	***1.29***	*1.09*	*0.97***
verbs / sentence	*2.45*	*2.92***	*2.67*	***2.72****	*2.72**	*2.47*
adj. and adv. / sentence	*2.77*	*2.91*	*2.51*	***2.69***	*2.95*	*2.5*
1st-person pron. / s-ce	*0.11*	*0.01**	*0.08*	***0.06***	*0.02**	*0.02**
proper nouns / sentence	*0.99*	*0.51*	*0.09****	***0.32******	*0.15****	*0.23****
hapax legomena	25.69	32.33	20.61	**19.3**	27.25	25.05
concreteness	2.48	2.42	2.46	**2.44**	2.37	2.4
pronouns / sentence	*0.95*	*0.64*	*0.87*	***0.88***	*0.66*	*0.73*
anaphora words / s-ce	*10.28*	*9.93*	*9.2*	***9.46***	*11.95*	*12.72*
cos. distance btwn s-ces	*0.12*	*0.1**	*0.18****	***0.15******	*0.14****	*0.14****
Euclidean distance from ref.	-	5.46	4.99	**3.82**	5.65	4.86

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 'Mistral, few-shot' corpus is represented in **bold** to denote its highest global closeness to the reference.

istics, please refer to Appendix 5: Distribution of Pedagogical Qualities per Corpus.

The corpus that is rated highest in the annotation process is 'Gemini, zero-shot', followed by 'Gemini, few-shot', 'Mistral, zero-shot', 'Mistral, few-shot' and 'Web-crawled' (see Table 2). The performance gap is largest between the web-crawled corpus (rated worst) and the second worst corpus, 'Mistral, few-shot', whilst the LLM-generated corpora exhibit higher similarity to one another. The figure in Appendix 5: Distribution of Pedagogical Qualities per Corpus clearly shows that the 'Web-crawled' corpus is the most frequent one to not receive the total number of point for all three investigated categories.

Interestingly, both corpora derived in zero-shot settings are rated more highly than their few-shot counterparts. The 'Gemini, few-shot' corpus is associated with the highest percentage of full points (71% of all contexts), followed by 'Gemini, zero-shot' (69%), 'Mistral, zero-shot' (61%), 'Mistral, few-shot' (60%) and 'Web-crawled' (29%). When the 'in-domain' characteristic is regarded in isolation, 'Gemini, few-shot' performs highest (by a small margin), and the rest of the classification remains the same. In turn, 'Mistral, few-shot' performs slightly better than 'Mistral, zero-shot' in relation to the 'ready to use' characteristic.This is also the characteristic for which the models shows largest variance in terms of the attribution of the highest number of points (see Appendix 5: Distribution of Pedagogical Qualities per Corpus).

In the free text notes, Mistral-generated text was surprisingly judged to have negative qualities that were explicitly addressed during the generation and filtering process: contexts were judged as too long in 8 cases in the zero-shot setting and 5 in the few-shot setting, a definition or explanation was provided instead of or along with the context (6 vs 2 instances), the pronoun 'I' was mentioned to have been used extensively (in 4 vs 2 examples), and the target word was said to have been closely repeated in one example (in the zero-shot setting). Therefore, the robustness of the applied filters should be examined. Other problems linked with examples generated by the model include lack of clarity (4 vs 1 instance), lack of informativeness (3 instances in the zero-shot setting), scientifically unsound text (3 instances in the zero-shot setting) and different meanings of the target word addressed (2 instances in the zero-shot setting). Perceivably fewer problems are noted in relation to the few-shot setting. In contrast, the issues noted in relation to Gemini-generated text,

while smaller in number, are not clearly reduced by way of the few-shot setting. Some contexts are judged to be too long (2 vs 5 instances), too generic (4 vs 4 instances) or unclear (2 vs 1 instances). Also, definitions or explanations were featured (2 vs 2 instances), and target words were used with a different meaning to the intended one (1 vs 3 instances). Finally, web-crawled examples were criticised for including quotations (4 instances), containing textual processing mistakes (2 instances) and being unclear (2 instances).

5 Discussion

Human evaluation rates the 'Gemini, zero-shot' corpus highest, while automatic comparison ranks 'Mistral, few-shot' first. In the case of Mistral, the few-shot setting seems to be efficient in reducing problems that make contexts not directly applicable in a classroom setting. Thus, the different corpora and, by definition, the derivation methods behind them are associated with different qualities and drawbacks.

Table 3 shows a juxtaposition of the contexts derived through all five described methods for the same ESP vocabulary item. The only context that did not receive the maximal number of points in the annotation was the web-crawled one, which was evaluated as not being entirely ready to use. Possible reasons could be its beginning with 'and', instances of complex grammar ('and though', 'those cases that did occur'), and the use of the definite article ('the procedure') when the reference is unknown to the reader. The web-crawled context is the longest, the 'Gemini, few-shot' the shortest, and the other three display similar length (19-20 words), which is also equal or close to that of the reference context (20 words). In the 'Mistral, zero-shot', 'Gemini, zero-shot' and 'Gemini, few-shot' contexts, the target word appears very close to the sentence's beginning, which is not the case with the reference. One could assume, therefore, that the 'Mistral, few-shot' setting has benefited from the proposed professional examples. Another specificity in the latter is the presence of a named entity ('The Second Law of Motion'). In terms of qualitative characteristics, one can claim that the reference context is scientifically sound and can serve an interdisciplinary purpose, and the same can interestingly be said about the two zero-shot LLM settings, which offer surprisingly similar examples, implying at the same time that the models' training suffices for a pedagogically apt formulation and that high similarity of output can be expected in the absence of narrow prompts and provided examples.

On the first research question, comparing web retrieval and generative AI, we observed that LLMs, when instructed using relevant prompt engineering and filtering techniques, are capable of providing contexts for the practice of ESP vocabulary that are evaluated by teaching experts as more pedagogically sound than counterparts retrieved from a corpus of scientific articles. In addition, examples of use generated by LLMs tend to share more textual characteristics with the ones handcrafted by professionals. The second research question also receives a positive reply as a large number of automatically derived contexts (290 out of 500) score maximally in terms of their pedagogical qualities based on human evaluation. In particular, 435 contexts received the maximum Likert value for the 'ready to use' quality. Finally, no clear correlation can currently be established between automatically derived contexts' textual characteristics and their pedagogical qualities (research question 3), as the two methods led to fully different classifications of the derivation methods.

The presented experiments and analyses extend current findings pertaining to the ability of LLMs to generate pedagogical contexts for the learning of foreign language vocabulary (such as the ones exposed by Paddags et al. (2024)) through the exploration of the models' few-shot abilities and the juxtaposition of human-based (qualitative) and automatic (quantitative) evaluation.

6 Conclusion

In this study, we demonstrated that high-quality contexts for an ESP vocabulary list can be obtained through contemporary NLP methods, in particular via LLM-based generation with prompt engineering. A possible problem is the reduced range and standard deviations that are associated with the derived contexts' measurable textual characteristics, which in turn may relate to a limited textual variety. A simple mitigation method would be the application of a variety of LLMs and generation settings to different vocabulary items, as they show different degrees of variation and adaptability to instructions. Other future directions of improvement may include the further adaptation of textual filters, such as a mod-

Corpus	In-domain	Comprehensible	Ready to Use	Overall
Web-crawled	89.2%	88.4%	78.4%	85.33%
Mistral, zero-shot	**99.4%**	97.8%	87.0%	94.73%
Mistral, few-shot	97.0%	95.8%	87.6%	94.37%
Gemini, zero-shot	97.2%	**99.0%**	**94.4%**	**96.87%**
Gemini, few-shot	97.6%	98.6%	90.4%	95.53%

Table 2: Percentages given to the derived corpora based on the human annotation process (as a portion of the total number of points possible).

Corpus	Sample context	In-d.*	Compr.*	RTU*
Reference	The Second Law of Motion states that the **rate** of change of momentum is directly proportional to the force applied.	-	-	-
Web-crawled	And though the **rate** of deaths associated with the procedure remained statistically flat, those cases that did occur were found with older patients.	5	5	4
Mistral, zero-shot	The **rate** of photosynthesis in plants depends on many factors such as temperature, light intensity and carbon dioxide concentration.	5	5	5
Mistral, few-shot	When calculating population growth **rates**, scientists use statistics to estimate the number of births and deaths in a given region.	5	5	5
Gemini, zero-shot	The **rate** of photosynthesis is influenced by the intensity of light, the availability of carbon dioxide, and the temperature.	5	5	5
Gemini, few-shot	The **rate** at which the climate changes is affected by human activity.	5	5	5

Table 3: Contexts for the item 'rate' (CEFR level B1, domain 'science') from the reference corpus and all automatically derived ones as well as the points the latter received in the human annotation. * Rating criteria: In-d. = In-domain; Compr. = Comprehensible; RTU = Ready to use.

ification of the permitted sentence lengths and domain filters that go beyond a binary classification of scientific vs non-scientific sentences. Finally, we are planning to make available a user-friendly online interface that facilitates the automatic generation of contexts based on selected ESP vocabulary items by teachers and students.

References

S. Aktay and G. G. D. Uzunoglu. 2023. Chatgpt in education: General applications as a dialogue agent and its impact on interaction, motivation, confidence, and vocabulary acquisition. *TAY Journal*, 7(2):378–406.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuliya Ardasheva, Tingting Hao, and Xiaobin Zhang. 2019. Pedagogical implications of current SLA research for vocabulary skills. In Nihat Polat, Peter D. MacIntyre, and Tammy Gregersen, editors, *Research-driven pedagogy: Implications of L2A theory and research for the teaching of language skills*, pages 125–144. Routledge.

Henri Besse. 1981. The pedagogic authenticity of a text. In *The Teaching of Listening Comprehension. Papers presented at the Goethe Institut Colloquium held in Paris in 1979*, pages 20–29. British Council.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Scott A. Crossley, Max M. Louwerse, Philip M. McCarthy, and Danielle S. McNamara. 2007. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30.

Robert Godwin-Jones. 2018. Evolving views on vocabulary development. *Language Learning & Technology*, 22(3):1–19.

Bahman Gorjian, Seyed Rahim Moosavinia, Kamal Elahi Kavari, Parsa Asgari, and Abouzar Hydarei. 2011. The impact of asynchronous computer-assisted language learning approaches on english as a foreign language high and low achievers' vocabulary retention and recall. *Computer Assisted Language Learning*, 24(5):383–391.

Michael Heilman, Le Zhao, Juan Pino, and Maxine Eskenazi. 2008. Retrieval of reading materials for vocabulary and reading practice. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 80–88, Columbus, Ohio. Association for Computational Linguistics.

Thomas Huckin and James Coady. 1999. Incidental vocabulary acquisition in a second language: a review. *Studies in Second Language Acquisition*, 21(2):181–193.

Anealka Hussin, Yuen Fook Chan, and Zubaidah Alsree. 2010. Scientific structural changes within texts of adapted reading materials. *English Language Teaching*, 3.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 28(12):15873–15892.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Diane J. Litman. 2016. Natural language processing for enhancing teaching and learning. In *AAAI Conference on Artificial Intelligence*.

David Little, Seán Devitt, and David Singleton. 1989. *Learning Foreign Languages from Authentic Texts: Theory and Practice*. Authentik.

Mathieu Loiseau, Georges Antoniadis, and Claude Ponton. 2005. Pedagogical text indexation and exploitation for language learning. In *Third international conference on multimedia and information and communication technologies in education (mICTE2005)*, volume 3 of *Recent Research Developments in Learning Technologies*, pages 984–994, Seville, Spain. Formatex.

Yihan Lou. 2023. Exploring the application of ChatGPT to english teaching in a malaysian primary school. *Journal of Advanced Research in Education*, 2(4):47–54.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Los Angeles, California. Association for Computational Linguistics.

Ana M. Morais and Isabel P. Neves. 2010. Educational texts and contexts that work discussing the optimization of a model of pedagogic practice. In

Daniel Frandji and Philippe Vitale, editors, *Knowledge, Pedagogy and Society: International Perspectives on Basil Bernstein's Sociology of Education*, page 18. Routledge, London.

Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin, and Thomas François. 2024. LLM-generated contexts to practice specialised vocabulary: Corpus presentation and comparison. In *Proceedings of TALN 2024*, Toulouse, France.

Benjamin Paddags, Daniel Hershcovich, and Valkyrie Savage. 2024. Automated sentence generation for a spaced repetition software. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 351–364, Mexico City, Mexico. Association for Computational Linguistics.

Robert Gilmore Pontius Jr and Marco Millones. 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15):4407–4429.

Falcon Dario Restrepo Ramos. 2015. Incidental vocabulary learning in second language acquisition: A literature review. *Profile: Issues in Teachers' Professional Development*, 17(1):157–166.

Irina Rets and Jekaterina Rogaten. 2021. To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification. *Journal of Computer Assisted Learning*, 37(3):705–717.

Angeliki Salamoura and Nick Saville. 2010. Exemplifying the CEFR: criterial features of written learner English from the english profile programme. In Inge Bartning, Maisa Martin, and Ineke Vedder, editors, *Communicative proficiency and linguistic development: intersections between SLA and language testing research*, number 1 in Eurosla Monographs Series, pages 101–132. Eurosla.

Sarang Shaikh, Sule Yildirim Yayilgan, Blanka Klimova, and Marcel Pikhart. 2023. Assessing the usability of ChatGPT for formal English language learning. *European Journal of Investigation in Health, Psychology and Education*, 13(9):1937–1960.

Try Siregar and Widyastuti Purbani. 2024. Prominent linguistic features of pedagogical texts to provide consideration for authentic text simplification. *Studies in English Language and Education*, 11:321–342.

Saifon Songsiengchai, Bank-on Sereerat, and Wirot Watananimitgul. 2023. Leveraging artificial intelligence (ai): Chatgpt for effective english language learning among thai students. *English Language Teaching*, 16(11):1–68.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, ..., and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models.

Brian Tomlinson. 2012. Materials development for language learning and teaching. *Language Teaching*, 45(2):143–179.

Julio Christian Young and Makoto Shishido. 2023a. Evaluation of the potential usage of ChatGPT for providing easier reading materials for ESL students. In *Proceedings of EdMedia and Innovate Learning*, pages 155–162. AACE.

Julio Christian Young and Makoto Shishido. 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).

Luo Yunjiu, Wei Wei, and Ying Zheng. 2022. Artificial intelligence-generated and human expert-designed vocabulary tests: A comparative study. *SAGE Open*, 12(1):21582440221082130.

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:

Here is a sentence[23] at CEFR level {***level***} showing how you use the **{*pos* if verb/noun/adverb/adjective; else 'word' or 'expression'}** "**{*item*}**" in the domain of {***domain***} (**{*lower*[24]}-{*upper*}** words):

Few-shot setting, level B1:

Please provide an example at level **B1** showing how you use the **{*pos*}** '**{*item*}**' (**{*domain*}**). Please use between **{*lower*}** and **{*upper*}** words.

Examples:
the adjective 'scarce': "As the planet continues to warm, resources such as freshwater, land, and food are becoming increasingly scarce."
the noun 'poaching': "As rhino populations decline rapidly due to habitat loss and poaching, the challenges for conservationists to protect these endangered species have never been more important."
the noun 'rate': "The carbon cycle is a complex process, and changes in land use and deforestation can affect the rate at which carbon is exchanged between the atmosphere and terrestrial ecosystems."
the verb 'reclaim': "The Great Green Wall is both an initiative for ecological restoration, and part of the fight against hunger and food insecurity in Africa. In existence since 2007, the wall is above all part of an immense effort to reclaim land lost to desertification."
the noun 'strain': "Before an earthquake occurs, tectonic plates accumulate strain along fault lines, gradually building up stress until it is released in a sudden rupture."

Few-shot setting,, level B2:

Please provide an example at level **B2** showing how you use the **{*pos*}** '**{*item*}**' (**{*domain*}**). Please use between **{*lower*}** and **{*upper*}** words.

Examples:
the noun 'spore': "Why some mushrooms are bioluminescent remains uncertain, but a study using LED

[23]The reason for 'sentence' to be used rather than 'example', even though some of the gold standard 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 few-shot setting, for which the word 'exmaple' is used instead

[24]'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.

lights adds to the evidence they attract insects that help the fungus disperse its spores."
the adjective 'bulbous': "Most of the evidence comes from soil fungi, many of which spend much of their life cycle as microorganisms, but also produce the bulbous fruiting bodies we know as mushrooms, toadstools, bracket fungi and the like. These are easy enough to spot, so they are often used as surrogates for the state of forest biodiversity, especially of the underground mycorrhizae – fungi that form symbiotic relationships with tree roots, taking sugars and supplying plants with water and mineral nutrients in return."
the noun 'shrub': "More recently, botanists in Brazil discovered six previously unknown species of fungus growing on the leaves of a tropical shrub, Coussapoa floccosa, which until recently was thought to be extinct. If and when the last specimen dies, those fungi will disappear too."
the verb 'undergo': "Nearly three-quarters of hammer coral colonies annually alternate between male and female. They are the only animal species known to undergo this change on such a regular schedule."
the noun 'brood': "Two species of bird have been observed raising offspring together. Such cooperative breeding between different species has never been documented before, says Rosario Balestrieri at the Stazione Zoologica Anton Dohrn of Naples, Italy. "It is a very strange and rare situation, in which the brood is mixed between the two species," he says."

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 percentage 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* (excluding end-of-s-ce punct.) 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 (as per Brysbaert et al. (2014)'s list of 40k English lemmas) 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 % of anaphora-denoting words per sentence* *average/min/max/SD cosine distance between sentences*

Table 4: Description of the linguistic features used in corpus comparison. The features marked in *italics* are representative continuous ones used in filters at the automatic derivation of contexts.

Appendix 4: Detailed Results of the Comparison between Corpora based on Textual Features

Entire Sample

Feature	Reference	Web-Crawled	Mistral: zero-shot	Mistral: few-shot	Gemini: zero-shot	Gemini: few-shot
Total # examples in sample	100	100	100	**100**	100	100
Total # words in sample	3787	2267	2823	**4091**	2345	2638
Avg. # words / s-ce	*13.33*	*11.11****	*11.67****	***13.73***	*11.17****	*11.32****
Min.	4	1	9	**2**	10	8
Max.	55	33	34	**44**	32	34
SD	8.48	4.82	5.08	**5.79**	5.34	6.08
Avg. # syllables / s-ce	*20.76*	*18.25****	*19.99**	***22.54****	*19.89****	*17.94*
Min.	6	1	14	**4**	15	14
Max.	85	60	63	**64**	62	56
SD	14.29	9.39	9.53	**9.86**	10.53	10.05
Avg. # letters / word	*5.2*	*5.37*	*5.57****	***5.4****	*5.84****	*5.21*
Min.	1	1	1	**1**	1	1
Max.	18	23	19	**16**	22	17
SD	2.8	3.02	3.06	**2.93**	3.2	2.87
Avg. # syllables / word	*1.56*	*1.64****	*1.71****	***1.64*****	*1.78****	*1.58*
Min.	0	0	1	**1**	1	0
Max.	7	9	6	**5**	6	6
SD	0.88	0.99	1.0	**0.93**	1.04	0.92
Avg. # noun phrases / s-ce	*5.76*	*6.34**	*6.08*	***6.29****	*6.26**	*5.68*
Min.	1	0	3	**1**	2	2
Max.	16	11	11	**14**	11	11
SD	2.75	1.95	1.85	**2.08**	1.93	1.96
Avg. % non-stem words / s-ce	*31.91*	*34.2*	*38.06****	***35.2****	*40.09****	*33.28*
Min.	5.88	0.0	11.11	**7.14**	17.39	7.14
Max.	61.54	63.64	66.67	**72.22**	69.23	54.55
SD	10.69	11.26	9.28	**10.42**	10.1	9.07
% s-ces ending in "?"	0.54	0.0	0.0	**0.51**	0.0	0.0
% s-ces ending in "!"	0.54	0.0	0.0	**0.0**	0.0	0.0
Avg. # punct. signs / s-ce	*1.51*	*0.98**	*1.06***	***1.29***	*1.09*	*0.97***
Min.	0	0	0	**0**	0	0
Max.	6	2	4	**6**	4	4
SD	0.54	0.35	0.43	**0.49**	0.4	0.4
Morphological richness	0.02	0.02	0.02	**0.02**	0.02	0.02
Avg. # verbs / s-ce	*2.45*	*2.92****	*2.67*	***2.72****	*2.72**	*2.47*
Min.	0	0	0	**0**	0	0
Max.	7	5	8	**7**	6	6
SD	1.51	1.04	1.38	**1.36**	1.13	1.17
Avg. # adj. and adv. / s-ce	*2.77*	*2.91*	*2.51*	***2.69***	*2.95*	*2.5*
Min.	0	0	0	**0**	0	0
Max.	8	6	7	**7**	7	7
SD	1.83	1.56	1.47	**1.49**	1.57	1.5
Avg. # 1st-person pron. / s-ce	*0.11*	*0.01**	*0.08*	***0.06***	*0.02**	*0.02**
Min.	0	0	0	**0**	0	0
Max.	2	1	1	**1**	2	1
SD	0.43	0.1	0.28	**0.23**	0.19	0.12
Avg. # proper nouns / s-ce	*0.99*	*0.51*	*0.09****	***0.32*****	*0.15****	*0.23****
Min.	0	0	0	**0**	0	0
Max.	11	4	2	**6**	2	3
SD	1.96	0.8	0.33	**0.87**	0.45	0.61

% words not in Dale-Chall list	44.71	46.18	46.09	**45.61**	50.36	43.48
% hapax legomena	25.69	32.33	20.61	**19.3**	27.25	25.05
Type-to-token ratio (words)	0.37	0.45	0.32	**0.31**	0.39	0.36
Type-to-token ratio (lemmas)	0.35	0.42	0.3	**0.29**	0.37	0.34
Average concreteness	2.48	2.42	2.46	**2.44**	2.37	2.4
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, and, in, a, for, is, that, are
Avg # pron. / s-ce	*0.95*	*0.64*	*0.87*	***0.88***	*0.66*	*0.73*
Min.	0	0	0	**0**	0	0
Max.	4	3	3	**3**	3	3
SD	1.1	0.7	0.79	**0.94**	0.75	0.76
Avg.% anaph. words / s-ce	*10.28*	*9.93*	*9.2*	***9.46***	*11.95*	*12.72*
Min.	0.0	0.0	0.0	**0.0**	3.23	0.0
Max.	27.27	22.73	23.81	**30.77**	25.0	25.0
SD	6.08	5.77	5.8	**5.99**	5.52	5.51
Avg.cos.d-ce btwn s-ces	*0.12*	*0.10**	*0.18****	***0.15******	*0.14****	*0.14****
Min.	-0.18	-0.19	-0.17	**-0.17**	-0.20	-0.20
Max.	0.70	1.0	0.80	**0.79**	0.84	0.84
SD	0.12	0.11	0.16	**0.14**	0.15	0.13

Per Level: B1 (domain 'Agronomy')

Feature	Reference	Web-Crawled	Mistral: zero-shot	Mistral: few-shot	Gemini: zero-shot	Gemini: few-shot
Total # examples in sample	56	56	56	56	56	56
Total # words in sample	1382	1179	1030	1581	971	892
Avg. # words / s-ce	*10.63*	*10.34****	*8.8**	*11.21*	*8.67****	*7.82****
Min.	4	1	11	2	10	8
Max.	41	27	26	31	26	25
SD	7.49	4.63	3.34	5.06	3.36	4.32
Avg. # syllables / s-ce	*16.65*	*16.88****	*14.93*	*18.57*	*15.36*	*12.45**
Min.	6	1	14	4	15	14
Max.	76	60	43	54	48	44
SD	12.69	9.34	6.47	8.58	7.22	7.54
Avg. # letters / word	*5.19*	*5.34*	*5.41*	*5.4*	*5.8****	*5.26*
Min.	1	1	1	1	1	1
Max.	16	23	15	15	22	17
SD	2.82	3.11	3.01	2.98	3.25	2.85
Avg. # syllables / word	*1.57*	*1.63*	*1.7***	*1.66**	*1.77****	*1.59*
Min.	0	0	1	1	1	1
Max.	7	9	5	5	6	6
SD	0.89	1.01	1.0	0.94	1.04	0.91
Avg. # noun phrases / s-ce	*5.08*	*5.75**	*5.15*	*5.62**	*5.05*	*4.41*
Min.	1	0	3	1	2	2
Max.	12	10	9	9	9	8
SD	2.38	1.73	1.28	1.65	1.41	1.24
Avg. % non-stem words / s-ce	*33.07*	*34.38*	*37.04**	*35.49*	*39.57****	*34.56*
Min.	10.0	0.0	16.67	14.29	17.39	15.79
Max.	61.54	61.54	56.25	53.85	66.67	54.55
SD	10.2	10.69	8.73	9.2	9.89	9.54
% s-ces ending in "?"	0.0	0.0	0.0	1.16	0.0	0.0
% s-ces ending in "!"	1.33	0.0	0.0	0.0	0.0	0.0
Avg. # punct. signs / s-ce	*1.16*	*0.86*	*0.57**	*1.01*	*0.77*	*0.55**
Min.	0	0	0	0	0	0
Max.	6	2	3	3	2	2
SD	0.49	0.32	0.28	0.38	0.29	0.25
Morphological richness	0.02	0.01	0.02	0.02	0.02	0.02
Avg. # verbs / s-ce	*2.15*	*2.67***	*2.2*	*2.26*	*2.2*	*1.97*
Min.	0	0	1	0	0	0
Max.	6	4	4	5	4	4
SD	1.33	1.06	0.96	1.08	0.96	1.03
Avg. # adj. and adv. / s-ce	*2.48*	*2.75*	*2.15*	*2.43*	*2.27*	*2.09*
Min.	0	0	0	0	0	0
Max.	6	6	6	6	5	5
SD	1.56	1.67	1.44	1.39	1.14	1.22
Avg. # 1st-person pron. / s-ce	*0.15*	*0.02*	*0.15*	*0.13*	*0.0**	*0.03*
Min.	0	0	0	0	0	0
Max.	2	1	1	1	0	1
SD	0.51	0.13	0.36	0.34	0.0	0.18
Avg. # proper nouns / s-ce	*0.77*	*0.49*	*0.15****	*0.17****	*0.14****	*0.19***
Min.	0	0	0	0	0	0
Max.	6	3	2	2	2	3
SD	1.28	0.73	0.4	0.47	0.4	0.51

% words not in Dale-Chall list	44.21	44.7	43.79	45.86	49.02	43.83
% hapax legomena	34.9	38.59	32.06	29.77	38.35	37.51
Type-to-token ratio (words)	0.46	0.5	0.44	0.42	0.49	0.48
Type-to-token ratio (lemmas)	0.45	0.49	0.42	0.4	0.47	0.46
Average concreteness	2.5	2.37	2.4	2.39	2.33	2.43
10 most frequent words (excl. stop words)	climate, water, change, earth, new, temperature, world, plants, two, greenhouse	development, may, next, many, european, climate, resources, would, human, possible	scientists, temperature, climate, change, growth, chemical, used, new, effects, plant	scientists, change, use, due, climate, science, world, around, universe, new	scientists, research, experiment, new, study, researchers, temperature, significant, scientist, effects	scientists, climate, due, change, study, used, light, energy, earth, human
10 most frequent words (incl. stop words)	the, of, to, in, and, is, a, on, are, that	the, to, of, and, a, in, that, is, will, be	the, to, of, and, in, a, that, scientists, for, can	the, to, of, and, in, is, a, that, for, scientists	the, of, to, and, in, a, that, is, scientists, can	the, of, to, in, is, and, scientists, a, can, are
Avg # pron. / s-ce	*0.8*	*0.61*	*0.8*	*0.86*	*0.52*	*0.55*
Min.	0	0	0	0	0	0
Max.	4	2	3	3	2	2
SD	1.1	0.59	0.79	0.84	0.6	0.6
Avg.% anaph. words / s-ce	*10.01*	*11.34**	*11.42*	*9.7*	*13.65**	*13.05*
Min.	0.0	0.0	3.85	0.0	4.35	4.55
Max.	27.27	22.73	22.22	30.77	25.0	25.0
SD	6.86	5.4	5.38	5.65	5.99	5.25
Avg.cos.d-ce btwn s-ces	*0.118*	*0.115***	*0.151****	*0.138****	*0.136****	*0.137****
Min.	-0.185	-0.145	-0.166	-0.167	-0.171	-0.200
Max.	0.701	1.000	0.765	0.751	0.584	0.758
SD	0.126	0.105	0.129	0.125	0.125	0.129

Per Level: B2 (domain 'Science')

Feature	Reference	Web-Crawled	Mistral: zero-shot	Mistral: few-shot	Gemini: zero-shot	Gemini: few-shot
Total # examples in sample	44	44	44	44	44	44
Total # words in sample	2405	1088	1793	2510	1374	1746
Avg. # words / s-ce	*15.62*	*12.09****	*14.34****	*15.99*	*14.02****	*14.67****
Min.	5	15	9	8	18	12
Max.	55	33	34	44	32	34
SD	8.85	4.09	5.02	5.74	3.59	4.9
Avg. # syllables / s-ce	*24.23*	*19.98****	*24.73***	*26.11**	*25.06****	*23.19**
Min.	10	26	16	14	29	17
Max.	85	55	63	64	62	56
SD	15.0	7.89	9.46	9.98	7.8	8.31
Avg. # letters / word	*5.21*	*5.39*	*5.67****	*5.41**	*5.87****	*5.18*
Min.	1	1	1	1	1	1
Max.	18	15	19	16	17	16
SD	2.79	2.94	3.09	2.91	3.16	2.89
Avg. # syllables / word	*1.55*	*1.65****	*1.72****	*1.63**	*1.79****	*1.58*
Min.	0	0	1	1	1	0
Max.	6	6	6	5	6	5
SD	0.87	0.96	0.99	0.92	1.04	0.93
Avg. # noun phrases / s-ce	*6.22*	*7.09**	*6.79*	*6.8*	*7.52****	*6.65*
Min.	1	3	3	2	4	2
Max.	16	11	11	14	11	11
SD	2.89	1.96	1.91	2.23	1.56	1.85
Avg. % non-stem words / s-ce	*31.25*	*34.01*	*38.63****	*35.01**	*40.47****	*32.62*
Min.	5.88	9.09	11.11	7.14	25.0	7.14
Max.	58.06	63.64	66.67	72.22	69.23	53.85
SD	10.97	12.09	9.65	11.29	10.36	8.69
% s-ces ending in "?"	0.91	0.0	0.0	0.0	0.0	0.0
% s-ces ending in "!"	0.0	0.0	0.0	0.0	0.0	0.0
Avg. # punct. signs / s-ce	*1.75*	*1.14**	*1.42*	*1.51*	*1.43*	*1.29**
Min.	0	0	0	0	0	0
Max.	4	2	4	6	4	4
SD	0.57	0.39	0.51	0.55	0.49	0.48
Morphological richness	0.02	0.02	0.02	0.02	0.02	0.02
Avg. # verbs / s-ce	*2.66*	*3.25***	*3.02*	*3.07**	*3.26***	*2.85*
Min.	0	1	0	0	1	0
Max.	7	5	8	7	6	6
SD	1.59	0.92	1.54	1.44	1.05	1.14
Avg. # adj. and adv. / s-ce	*2.96*	*3.11*	*2.78*	*2.88*	*3.65***	*2.83*
Min.	0	1	0	0	0	0
Max.	8	5	7	7	7	7
SD	1.98	1.4	1.44	1.53	1.65	1.61
Avg. # 1st-person pron. / s-ce	*0.09*	*0.0*	*0.04*	*0.0***	*0.04*	*0.0**
Min.	0	0	0	0	0	0
Max.	2	0	1	0	2	0
SD	0.37	0.0	0.19	0.0	0.27	0.0
Avg. # proper nouns / s-ce	*1.14*	*0.55*	*0.05****	*0.43**	*0.17***	*0.25***
Min.	0	0	0	0	0	0
Max.	11	4	2	6	2	3
SD	2.31	0.87	0.27	1.07	0.5	0.68

% words not in Dale-Chall list	44.99	47.79	47.41	45.46	51.31	43.3
% hapax legomena	29.88	40.86	20.83	22.35	29.53	28.07
Type-to-token ratio (words)	0.42	0.53	0.33	0.34	0.41	0.39
Type-to-token ratio (lemmas)	0.4	0.51	0.31	0.32	0.39	0.37
Average concreteness	2.47	2.47	2.5	2.48	2.4	2.38
10 most frequent words (excl. stop words)	water, species, trees, carbon, could, wild, forests, researchers, reef, world	could, development, climate, change, international, benefits, health, people, responsibility, study	soil, crop, agriculture, farmers, crops, practices, use, sustainable, farming, growth	soil, crop, farmers, crops, water, use, growth, used, agriculture, levels	crop, soil, water, agricultural, farmers, growth, drought, conditions, practices, yields	species, soil, plant, plants, fungi, new, crop, crops, water, nutrients
10 most frequent words (incl. stop words)	the, of, and, to, in, a, is, that, are, for	to, the, and, of, a, in, with, that, for, or	and, the, to, of, soil, can, in, crop, a, agriculture	and, the, to, of, soil, can, in, crop, a, for	the, of, and, to, for, in, a, crop, as, their	the, of, and, to, a, in, that, is, for, as
Avg # pron. / s-ce	*1.05*	*0.68*	*0.91*	*0.9*	*0.81*	*0.87*
Min.	0	0	0	0	0	0
Max.	4	3	3	3	3	3
SD	1.1	0.83	0.79	1.0	0.85	0.84
Avg.% anaph. words / s-ce	*10.47*	*8.1*	*7.53*	*9.27*	*10.19*	*12.46***
Min.	0.0	0.0	0.0	0.0	3.23	0.0
Max.	25.0	20.69	23.81	27.78	20.69	25.0
SD	5.52	5.78	5.57	6.26	4.38	5.72
Avg.cos.d-ce btwn s-ces	*0.14*	*0.11****	*0.38****	*0.24****	*0.32****	*0.24****
Min.	-0.15	-0.17	-0.01	-0.10	-0.10	-0.08
Max.	0.63	1.00	0.80	0.79	0.84	0.84
SD	0.12	0.12	0.14	0.16	0.14	0.13

Features in *italics* have been tested for statistical significance, and the extent of the significance is marked with *, ** and *** from lowest to highest. The few-shot Mistral corpus is marked with **bold** when the entire corpora are considered to denote its highest global similarity to the reference corpus.

Appendix 5: Distribution of Pedagogical Qualities per Corpus