Leading by Example: The Use of Generative Artificial Intelligence to Create Pedagogically Suitable Example Sentences

Jasper Degraeuwe and **Patrick Goethals**
LT^3 / MULTIPLES
Ghent University
Belgium
Jasper.Degraeuwe@UGent.be, Patrick.Goethals@UGent.be

Abstract

Studies on second language acquisition have argued in favour of practising vocabulary in authentic contexts. After the tradition of obtaining these usage examples by "invention" (i.e. language experts creating examples based on their intuitions) was superseded by corpus-based approaches (i.e. using dedicated tools to select examples from corpora), the rise of large language models led to a third possible "data source": Generative Artificial Intelligence (GenAI). This paper aims to assess GenAI-based examples in terms of their pedagogical suitability by conducting an experiment in which second language (L2) learners compare GenAI-based examples to corpus-based ones, for L2 Spanish. The study shows that L2 learners find GenAI-based sentences more suitable than corpus-based sentences, with – on a total of 400 pairwise comparisons – 265 artificial examples being found most suitable by all learners (compared to 10 corpus-based examples). The prompt type (different zero-shot and few-shot prompts were designed) did not have a noticeable impact on the results. Importantly, the GenAI approach also yielded a number of unsuitable example sentences, leading us to conclude that a "hybrid" method which takes authentic corpus-based examples as its starting point and employs GenAI models to rewrite the examples might combine the best of both worlds.

1 Introduction

Although vocabulary items can be learnt in isolation (e.g., through flash cards; Nation, 2022), providing in-context usage examples of vocabulary items strengthens word form - word meaning associations (Laufer and Shmueli, 1997) and has shown to foster both language comprehension and production (Frankenberg-Garcia, 2012, 2014). As a result, example sentences are often used in vocabulary lists, learners' dictionaries, and grammar sections as a means to illustrate the usage(s) of vocabulary items and grammatical patterns. Some types of materials even depend entirely on the presence of example sentences, such as fill-in-the-blanks and in-context translation exercises.

To obtain example sentences, linguistic disciplines have a long tradition of using intuited/invented examples (IEs) created by language experts such as lexicographers and teachers (Cook, 2001; Laufer, 1992; Stefanowitsch, 2020). The underlying idea is that their advanced linguistic competence allows them to formulate well-formed, relevant, and grammatically correct sentences. However, the last decades witnessed an increased interest in the selection of example sentences from digital(ised) native (L1) corpora, first manually and later following (semi-)automatic selection procedures (Frankenberg-Garcia et al., 2021). Even though well-designed IEs can have pedagogical value (Cook, 2001), carefully selected corpus examples can be considered more authentic, reliable, and valid expressions of language (Firth, 1968; Stefanowitsch, 2020). Moreover, thanks to the continued improvements made to the tools and techniques used for corpus processing and consultation, performing corpus queries to extract sentences that should meet a given set of criteria has become highly efficient.

Recently, major developments in the field of Generative Artificial Intelligence (GenAI) uncovered another pathway to obtain example sentences: based on a prompt specifying the desired criteria, GenAI systems can be asked to output a series of – according to the model – suitable usage examples. Although the artificial way in which they are conceived bears some resemblance with IEs, these examples can also be said to have a corpus-based touch, since the GenAI tools that produce them are trained on (extremely large) collections of text.

In the present paper, we present an experiment in which second language (L2) learners of Spanish compare example sentences selected following a *corpus-based method* to examples created following a *GenAI-based method*. In doing so, we aim to make a contribution to the assessment of the pedagogical usability and validity of artificially generated learning materials. The paper is structured as follows: after providing an overview of the related research in Section 2, we describe the methodology (Section 3) and elaborate on the results (Section 4). A discussion of those results is presented in Section 5. Finally, Section 6 includes a conclusion together with possible directions for future research.

2 Related Research

Broadly speaking, the criteria which define a "good" example can be categorised as either form-related or content-related. The former type refers to grammatical soundness and straightforward superficial properties such as a capitalised first letter and a punctuation mark at the end of the sentence. Content-related criteria, on the other hand, encompass features such as naturalness (i.e. containing formulations which can also be encountered in real-life language use), context independence, intelligibility (often captured in terms of sentence length and number of difficult words), typicality (i.e. containing collocations or colligations), and informativeness (i.e. containing clues which help understand the meaning of the target item).

The definition of sentence selection criteria has been considered from both a pedagogical (Pilán et al., 2016) and a lexicographic point of view (Atkins and Rundell, 2008). Although many criteria apply to both of them, the two perspectives also exhibit differences. With regard to the intelligibility criterion, lexicographic resources tend to prefer short sentences, while language learning resources are considerably more tolerant towards long sentences, as exposing learners to more (relevant) context can be beneficial for the learning process (Kosem et al., 2019). Secondly, in a language learning setting, the criteria of informativeness and typicality are often isolated and linked to, respectively, the concepts of "decoding" (i.e. aimed at fostering comprehension) and "encoding" (i.e. aimed at fostering production). As these concepts reflect two very distinct aspects of language learning, the example selection methods used to create language learning resources often focus on only one of these two criteria, instead of looking for sentences incorporating both (Frankenberg-Garcia, 2014). Finally, selecting sentences for pedagogical purposes also requires assessing a sentence's complexity in terms of learner proficiency levels and adapting the selection accordingly, as there exist considerable differences between the language knowledge of beginning, intermediate, and advanced learners.

2.1 Corpus-based Examples

Finding its origins in the grammar-translation method of the mid-19th century, invented examples (IEs) have long been the primary source for presenting new words or exemplifying linguistic phenomena of a lexical (i.e. collocations) or grammatical (i.e. colligations) nature (Cook, 2001). In essence, IEs are concocted by experts (e.g., L2 teachers or lexicographers) and rely on the intuitions these experts have about the usage of the word/pattern to be presented/exemplified. Towards the end of the 20th century, however, the rise of online accessible corpora together with advances in the technological means to process and consult them opened new horizons in the selection/creation of examples. The COBUILD initiative (Sinclair, 1987), for example, radically rejected the use of IEs and only used unaltered corpus examples in its resources.

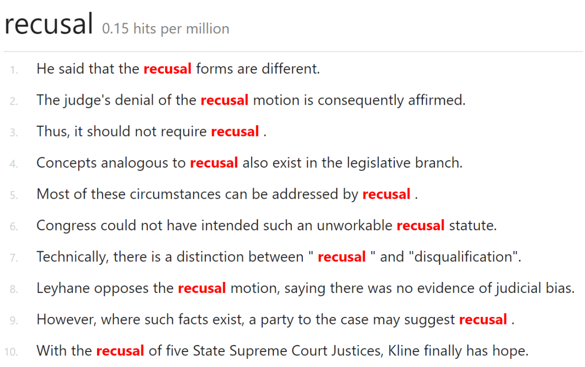
Importantly, much of this research into corpus-based example selection methods originated from lexicographic motives, which – as mentioned earlier – do not necessarily include pedagogical considerations. Yet, many lexicographic methods were (and still are) also used for pedagogical purposes (Kosem et al., 2019). One of those methods is GDEX (Good Dictionary EXamples; Kilgarriff et al., 2008), which marked a major milestone in the field of corpus-based example selection. In brief, the method takes as input a list of corpus concordances for a given target item and returns a ranked version of that list. The main particularities of GDEX are the overall scoring algorithm with adjustable parameters (a so-called "GDEX configuration") and the "second collocate" classifier that prioritises sentences containing the most typical collocates of a given collocation. Moreover, as the adjustable parameters allow users to tailor the sentence selection criteria to their specific needs, the need for posterior manual revisions also decreases.

As mentioned above, GDEX is – despite its lexicographic origins – widely applied in language learning contexts as well (Kallas et al., 2015; Smith et al., 2010). The SKELL tool, for example, employs GDEX to retrieve the most useful examples for language learners from large corpora and return them as a ranked list (see Figure 1). Nevertheless, extra curation is still required when selecting examples from GDEX-based concordances, particularly when priority has to be given to specific collocation or colligation patterns (Frankenberg-Garcia et al., 2021).

Regarding the (limited) research dedicated to corpus-based sentence selection specifically for language learning purposes, a first important study to highlight is that on HitEx (Pilán et al., 2016), a sentence selection framework for L2 Swedish. Combining both rule-based and machine learning-based components, the HitEx framework pays special attention to linguistic complexity and independence from the surrounding corpus sentences, but also takes into account well-formedness and a series of structural criteria (e.g., presence of modal verbs and sentence length) and lexical criteria (e.g., word frequency and presence of proper names). Next, Heck and Meurers (2022) developed an algorithm which can select suitable examples to be used as input for L2 English grammar exercises. Apart from offering different data sources to choose from (the web, precompiled corpora, or custom texts), the method also includes tailor-made selection criteria such as the presence of relative pronouns, extraposition, and preposition stranding.

2.2 GenAI-based Examples

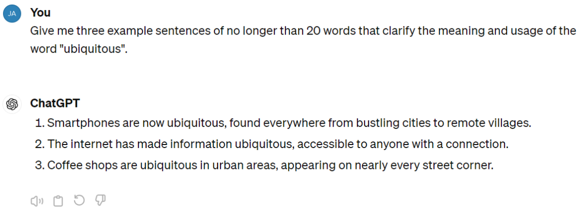
The process to obtain artificially generated example sentences is very straightforward: based on a natural language prompt as input, a GenAI model can be asked to return a series of sentences, without any specific prior training. Depending on the model's architecture, the prompt can be formulated as a *zero-shot learning* or *few-shot learning* phrase. As shown in Figure 2, zero-shot prompts can be written as if one is making a request/asking a question to a fellow human being. In this case, we simply ask the model for three sentences that have to meet a set of criteria (sentences cannot be longer than 20 words and have to clarify the meaning and usage of the target item). With few-shot prompting, the request/question is complemented by (or sometimes even replaced by) a limited number of examples the model can learn from, as illustrated in Figure 3. In this case, we just take the three sentences returned by the model for the zero-shot query, convert them into a structured format, and prompt the model to return the corresponding information for a new item. The underlying idea is that the model will "deduce" the desired characteristics from the examples (e.g., the sentence length) and use this information when generating the response for the new items.



recusal 0.15 hits per million

1. He said that the **recusal** forms are different.
2. The judge's denial of the **recusal** motion is consequently affirmed.
3. Thus, it should not require **recusal** .
4. Concepts analogous to **recusal** also exist in the legislative branch.
5. Most of these circumstances can be addressed by **recusal** .
6. Congress could not have intended such an unworkable **recusal** statute.
7. Technically, there is a distinction between " **recusal** " and "disqualification".
8. Leyhane opposes the **recusal** motion, saying there was no evidence of judicial bias.
9. However, where such facts exist, a party to the case may suggest **recusal** .
10. With the **recusal** of five State Supreme Court Justices, Kline finally has hope.

Figure 1: SKELL output for *recusal*. Date of performing query: 6 June 2024.



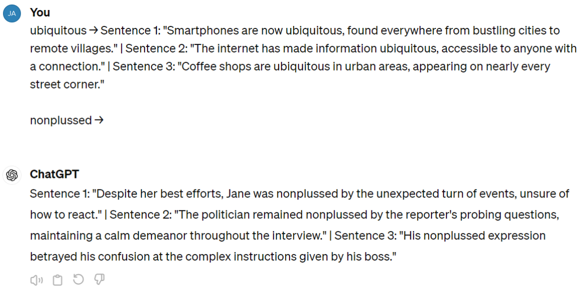
You

Give me three example sentences of no longer than 20 words that clarify the meaning and usage of the word "ubiquitous".

ChatGPT

1. Smartphones are now ubiquitous, found everywhere from bustling cities to remote villages.
2. The internet has made information ubiquitous, accessible to anyone with a connection.
3. Coffee shops are ubiquitous in urban areas, appearing on nearly every street corner.

Figure 2: Artificially generated example sentences by means of zero-shot learning (i.e. a simple instruction/question) as prompting technique. Model: OpenAI's GPT-3.5 (accessed through ChatGPT interface). Date of performing prompt: 6 May 2024.



You

ubiquitous → Sentence 1: "Smartphones are now ubiquitous, found everywhere from bustling cities to remote villages." | Sentence 2: "The internet has made information ubiquitous, accessible to anyone with a connection." | Sentence 3: "Coffee shops are ubiquitous in urban areas, appearing on nearly every street corner."

nonplussed →

ChatGPT

Sentence 1: "Despite her best efforts, Jane was nonplussed by the unexpected turn of events, unsure of how to react." | Sentence 2: "The politician remained nonplussed by the reporter's probing questions, maintaining a calm demeanor throughout the interview." | Sentence 3: "His nonplussed expression betrayed his confusion at the complex instructions given by his boss."

Figure 3: Artificially generated example sentences by means of few-shot learning as prompting technique. Model: OpenAI's GPT-3.5 (accessed through ChatGPT interface). Date of performing prompt: 6 May 2024.

The GenAI-driven creation of example sentences has been explored in the context of data-driven learning (Crosthwaite and Baisa, 2023) and as a means to clarify difficult words (Kohnke et al., 2023). A large-scale study which specifically assesses the pedagogical suitability of artificially generated example sentences has not yet been performed, a gap we aim to fill with this study. However, an important observation to make in the context of GenAI research is that the non-deterministic nature of the (online) models makes the research, per definition, irreplicable. Due to randomness being included in the generation process, GenAI models can produce different outputs at different times for the same input prompts[1]. Regular updates to existing models (e.g., of OpenAI's proprietary GPT-3.5) and launches of new models (e.g., OpenAI's GPT-4 and GPT-4o or Google's Gemini models) further complicate adequately assessing the pedagogical value of artificially generated sentences. Nevertheless, even given these methodological drawbacks, there is a growing consensus that scientific research is needed to explore the use of GenAI models for the creation of all kinds of L2 learning materials and to help shed light on the pedagogical suitability of this approach (Crosthwaite and Baisa, 2023; Caines et al., 2023).

3 Methodology

As mentioned in the introduction, our aim is to evaluate the pedagogical usability of artificially generated example sentences by comparing them to corpus-based sentences, which have become the standard approach for obtaining pedagogically suitable example sentences. To this end, we organise an experiment in which L2 Spanish learners compare corpus examples selected according to a dedicated sentence selection framework (Section 3.2) with examples generated by means of OpenAI's GPT-3.5 Turbo model, using different types of prompts (Section 3.3). In total, we recruit seven students from both beginner and advanced proficiency levels, all with Dutch as their L1 (see Section 3.4 for more details). For the former group, we envisage a general vocabulary learning course as the target setting; for the latter, we take a language for specific purposes course on legal vocabulary as our anchor point. The research questions we aim to answer are defined as follows:

1. Which source of example sentences is found most suitable by L2 learners: corpus-based or GenAI-based?
2. Which type of prompt used to query the GenAI model is found most suitable by L2 learners: zero shot (with varying degrees of specificity) or few shot?

3.1 Dataset

For each of the two target groups (beginner and advanced), we collect a set of 250 target items, which are selected based on their relevance and representativeness for the target setting defined above. For the beginner group, we take the first 150 nouns, 50 verbs, and 50 adjectives from the 1,001-2,000 frequency range in the Davies and Hayward Davies (2018) word list, excluding Spanish-Dutch cognates (e.g., ES *proyecto* - NL *project* - EN *project*). For the advanced group, we take a 25M specialised corpus containing newspaper articles on legal topics[2] as our starting point, rank all words in the corpus based on Odds Ratio as the keyness metric (Pojanapunya and Watson Todd, 2018; Gabrielatos, 2018) and select the first 150 nouns, 50 verbs, and 50 adjectives from the resulting list. Apart from cognates, we also exclude region-specific eponyms (e.g., *baltarismo*, which refers to the political movement named after the Galician politician José Manual Baltar) and derivations with *ex*, *sub*, and *vice* as the prefixes (e.g., *exdiputado*: 'former MP'; *subgobernador*: 'vice governor'; *vicepresidente*: 'vice president').

3.2 Corpus-based Examples

To obtain corpus-based sentences for the 500 target items, we develop a dedicated framework to select examples from corpora. Our framework – named SelEjemCor (**Sel**ección de **Ejem**plos de **Cor**pus) – builds on the work of Pilán et al. (2016), who developed the HitEx sentence selection framework for L2 Swedish (see also Section 2.1). In comparison to HitEx, our framework – the first of its kind for L2 Spanish – includes the integration of a tailor-made word difficulty classifier and the promotion of *typicality*

[1] Recently, many large language model providers added a "seed" parameter to their (API) interface, allowing developers to receive (mostly) consistent outputs. Yet, due to the inherent non-determinism of GenAI models, there will always be a small chance that responses differ even when a seeding parameter is specified.

[2] The corpus is available within the Spanish Corpus Annotation Project (SCAP; Goethals, 2018).

to being a main selection criterion as novel aspects. A comprehensive overview of all selection criteria included in the framework is presented in Appendix A. The Python implementation of the framework is made publicly available in a GitHub repository. To obtain the morphosyntactic information required for certain selection criteria (e.g., on part-of-speech tags, morphosyntactic features, and dependency relations), the Python module makes use of spaCy's automatic morphosyntactic analysis pipeline[3]. To render our framework as language-independent as possible, we use the morphosyntactic categories and labels proposed by the Universal Dependencies initiative (Nivre et al., 2016).

All Boolean criteria in SelEjemCor function as *filters* (i.e. if the criterion is not met, the sentence will be excluded from the selection), whereas all numerical criteria function as *rankers* (i.e. the closer the numerical value lies to the desired value, the higher the sentence will be ranked). For filters, criterion values can be set to either *True* (filter active, all sentences which do not pass the filter are excluded) or *None* (filter inactive). For rankers, values can be set to any numerical value (in which case the criterion will act as a threshold-based ranker, with all sentences obtaining a better value than the threshold being considered equally suitable), to *all* (in which case the selection algorithm will simply rank all sentences from highest to lowest value), or to *None* (ranker inactive). In the end, all sentences which have not been filtered out receive one single overall "goodness score", which corresponds to the average of all individual ranking positions.

We apply the SelEjemCor framework to a 7.5M corpus containing accessible reportages about tourist destinations[4] (for the 250 items in the beginner group) and to the abovementioned 25M specialised corpus containing newspaper articles on legal topics (for the 250 items in the advanced group). For each target item, we select the top-ranked sentence according to the selection algorithm explained above. The values set for the different selection criteria are included in Appendix A. For the advanced group, we make the values slightly more tolerant in terms of non-lemmatised tokens, modal verbs, word frequency, and out-of-vocabulary words.

3.3 GenAI-based Examples

To obtain the artificially generated sentences, we use OpenAI's GPT-3.5 Turbo model. We define four different prompt types and corresponding prompt texts to access the model, with the prompt texts also varying according to the target group. The prompts define both a *system role* (which specifies the *way* in which the model answers questions) and a *user role* (which specifies the *output* that should be returned). A short description of the prompt types is provided below (see Appendix B for the full overview):

1. **ZS-GEN** (zero-shot general): only the broad context (L2 learning setting; Spanish as target language; desired sentence length; sentence has to be usage example) is included in the prompt.

2. **ZS-GEN+AUD** (ZS-GEN plus target audience): apart from the broad context, also the target audience is specified in the prompt.

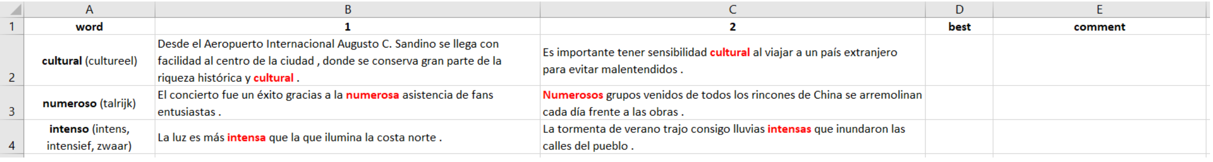
3. **ZS-GEN+AUD+CRIT** (ZS-GEN+AUD plus criteria): next to the broad context and the target audience, the prompt also includes the specific "goodness" criteria the output sentence should adhere to.

4. **FEWSHOT**: a limited number of suitable sentences (one sentence for each part of speech; with target words that do not occur in dataset) are provided in the prompt for the model to learn. The example words are selected from the 2,001-3,000 frequency range in Davies and Hayward Davies (2018) and the sentences are extracted from the Spanish Clave dictionary (González, 2012). The prompt also presents the broad context and differentiates between the two target audiences (see ZS-GEN and ZS-GEN+AUD above).

To enable the analysis at the layer of the prompt type, we randomly subdivide the 250 items in each group (beginner and advanced) into five subsets of 50 items (30 nouns, 10 verbs, and 10 adjectives). For the 50 items in the first subset (IDs 1 and 6), we generate an example sentence based on the ZS-GEN prompt type; for the second set (IDs 2 and

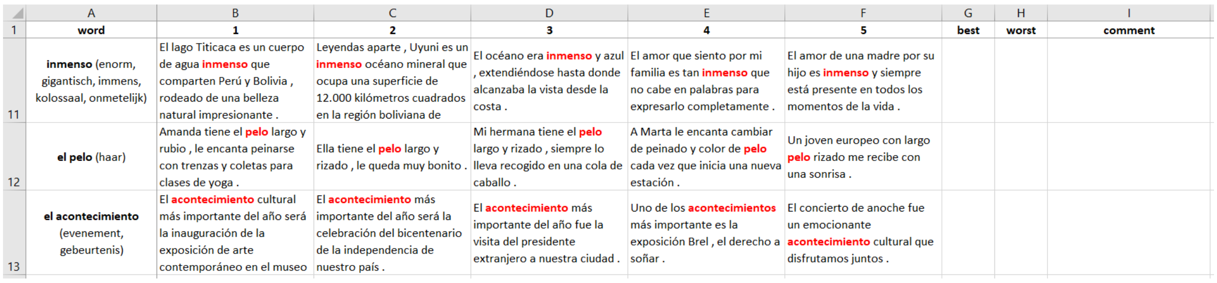
[3] Even though other NLP toolkits such as UDPipe and Stanza tend to perform (slightly) better at tagging and parsing natural text, spaCy's built-in large and Transformer-based models have shown to achieve near state-of-the-art performance with a significantly higher processing speed.

[4] Also compiled within SCAP.



	A	B	C	D	E
1	**word**	**1**	**2**	**best**	**comment**
2	**cultural** (cultureel)	Desde el Aeropuerto Internacional Augusto C. Sandino se llega con facilidad al centro de la ciudad , donde se conserva gran parte de la riqueza histórica y **cultural** .	Es importante tener sensibilidad **cultural** al viajar a un país extranjero para evitar malentendidos .		
3	**numeroso** (talrijk)	El concierto fue un éxito gracias a la **numerosa** asistencia de fans entusiastas .	**Numerosos** grupos venidos de todos los rincones de China se arremolinan cada día frente a las obras .		
4	**intenso** (intens, intensief, zwaar)	La luz es más **intensa** que la que ilumina la costa norte .	La tormenta de verano trajo consigo lluvias **intensas** que inundaron las calles del pueblo .		

Figure 4: Example of pairwise comparison between corpus-based and GenAI-based example sentences (in subset 1 to 4 and subset 6 to 9). The order in which the sentences are presented is randomised.



	A	B	C	D	E	F	G	H	I
1	**word**	**1**	**2**	**3**	**4**	**5**	**best**	**worst**	**comment**
11	**inmenso** (enorm, gigantisch, immens, kolossaal, onmetelijk)	El lago Titicaca es un cuerpo de agua **inmenso** que comparten Perú y Bolivia , rodeado de una belleza natural impresionante .	Leyendas aparte , Uyuni es un **inmenso** océano mineral que ocupa una superficie de 12.000 kilómetros cuadrados en la región boliviana de	El océano era **inmenso** y azul , extendiéndose hasta donde alcanzaba la vista desde la costa .	El amor que siento por mi familia es tan **inmenso** que no cabe en palabras para expresarlo completamente .	El amor de una madre por su hijo es **inmenso** y siempre está presente en todos los momentos de la vida .			
12	**el pelo** (haar)	Amanda tiene el **pelo** largo y rubio , le encanta peinarse con trenzas y coletas para clases de yoga .	Ella tiene el **pelo** largo y rizado , le queda muy bonito .	Mi hermana tiene el **pelo** largo y rizado , siempre lo lleva recogido en una cola de caballo .	A Marta le encanta cambiar de peinado y color de **pelo** cada vez que inicia una nueva estación .	Un joven europeo con largo **pelo** rizado me recibe con una sonrisa .			
13	**el acontecimiento** (evenement, gebeurtenis)	El **acontecimiento** cultural más importante del año será la inauguración de la exposición de arte contemporáneo en el museo	El **acontecimiento** más importante del año será la celebración del bicentenario de la independencia de nuestro país .	El **acontecimiento** más importante del año fue la visita del presidente extranjero a nuestra ciudad .	Uno de los **acontecimientos** más importante es la exposición Brel , el derecho a soñar .	El concierto de anoche fue un emocionante **acontecimiento** cultural que disfrutamos juntos .			

Figure 5: Example of BWS comparison between corpus-based and GenAI-based example sentences (in subset 5 and 10). The order in which the sentences are presented is randomised.

Prompt type	**Subset ID**	
	BEG	ADV
ZS-GEN	1	6
ZS-GEN+AUD	2	7
ZS-GEN+AUD+CRIT	3	8
FEWSHOT	4	9
ALL	5	10

Table 1: Overview of prompt types used to generate artificial example sentences. "BEG" stands for beginner, "ADV" for advanced.

7) based on ZS-GEN+AUD; for the third set (IDs 3 and 8) based on ZS-GEN+AUD+CRIT; and for the fourth set (IDs 4 and 9) based on FEWSHOT (see Table 1). For the 50 items in the fifth subset (IDs 5 and 10), we generate an artificial example sentence based on all four prompt types. Finally, a Dutch translation is added for all 500 target items in the dataset (see Table 2 for a dataset sample).

3.4 Evaluation Procedure

For each of the two target audiences (beginner and advanced), the first four subsets are used to perform pairwise comparisons between corpus-based sentences and artificially generated ones. As the artificial sentences are generated based on different prompts, comparing the results at subset level will also enable us to gain insights into the performance of each prompt type. The fifth subset is used to compare all five possible sentence sources (i.e. corpus-based and the four different GenAI prompts) at once in a best-worst scale (BWS) setup. The 250 beginner items are evaluated by three L2 Spanish learners (≈ B1 proficiency level, 19 years old, L1 Dutch), the 250 advanced items are assessed by four learners (≈ C1 proficiency level, 22-24 years old, L1 Dutch)[5].

Prior to starting the experiment, participants were given a written document including the instructions, which were discussed orally with one of the researchers involved in the study. In the pairwise comparisons, participants were asked to indicate the best sentence, as illustrated in Figure 4; in the BWS comparisons, they were asked to indicate both the best and the worst one, as illustrated in Figure 5. To make the term "best" as concrete as possible, the instructions stipulated that the participants should first check if the sentences complied with a series of criteria, which are explained below. Together, these descriptions reflect how the term "pedagogical suitability" as used in this paper should be interpreted.

- The sentence is not a definition. If it is, the participant should write "definition" in the

[5] All students are enrolled in the Applied Linguistics study career at Ghent University, Belgium. The Applied Linguistics curriculum stipulates – based on the CEFR scale – the minimal linguistic competences students should gain before they are admitted to the next year of the career. As a result, we can estimate the proficiency level of the learners based on the year they are enrolled in.

Item	POS	Value	ID	Corpus-based	GenAI-based
enemigo (‘vijand’)	NOUN	1,024	1	El coche es el **enemigo** público número uno: en Londres se aplica una tasa ambiental a los vehículos más contaminantes.	Durante la guerra, es importante reconocer quién es tu verdadero **enemigo** para poder luchar de manera estratégica y efectiva.
causar (‘veroorzaken’)	VERB	1,007	3	Los bares y restaurantes de madera **causan** una impresión de poblado tradicional.	El exceso de velocidad puede **causar** accidentes graves en la carretera.
político (‘politiek’)	ADJ	1,237	5	Los de los partidos **políticos** acompañan a sus votantes en la otra vida.	• Es importante estar informado sobre la situación político-social de un país para comprender su realidad y desarrollo. • El discurso **político** del presidente generó opiniones divididas entre la población. • La situación **política** en América Latina es muy complicada debido a diversos factores económicos y sociales. • El discurso **político** del presidente fue muy persuasivo y tuvo gran impacto en la opinión pública.
exacción (‘heffing’)	NOUN	75.8	7	La investigación le atribuye presuntos delitos de cohecho, prevaricación, blanqueo de capitales y fraude y **exacciones** ilegales.	La **exacción** de impuestos a menudo genera debate y controversia en la sociedad.
deslegitimar (‘delegitimeren’)	VERB	206	9	Los independientes, a su modo de ver, “**deslegitiman** y desnaturalizan la participación de los partidos”.	El periódico publicó un artículo que intentó **deslegitimar** las acusaciones contra el político.

Table 2: Dataset sample. “ID” refers to the subset ID. Values for the beginner group (subset 1-5) refer to the rank in Davies and Hayward Davies (2018); values for the advanced group (subset 6-10) refer to the Odds Ratio value.

“comment” column and annotate the other sentence as “best”.

- The sentence can be understood without any additional context (i.e. it is context-independent). If not, the participant should write “context-dependent” in the “comment” column and annotate the other sentence as “best”.
- The sentence does not contain words that are too difficult. If it does, the participant should write “too difficult” in the “comment” column and annotate the other sentence as “best”.

In case the example sentences adhered to all criteria, participants were instructed to indicate which sentence they found best (and worst in case of the BWS setup) based on their intuitions and needs as L2 learners. Regarding measures taken to arrive at qualitative annotations, we organised the first batch of ten annotations as an on-site session without any time constraints, allowing us to provide guidance and answer questions whenever necessary. The remaining annotations could be completed at home. For their annotation work, the participants also received a financial compensation, serving as an additional incentive for them to complete the classification task diligently.

Finally, we checked if the sentences complied with the following formal criteria:

- The target item occurs in the sentence. If not, we label the other sentence as “best”[6].
- The target item has the correct part of speech (POS). If not, we label the other sentence as “best”.
- The sentence is complete (i.e. it starts with capital letter and ends with punctuation mark). If not, we label the other sentence as “best”.

[6] Unless the target item does also not occur in that sentence, in which case we label both sentences as “N/A”.

4 Results

The results of the experiment have been summarised into a series of tables, listed below. The tables will be extensively referred to in our two main analyses: the comparison between corpus-based and GenAI-based as the source of the sentence (RQ1; Section 4.1) and the comparison between the different prompt types to generate the artificial example sentences (RQ2; Section 4.2).

- Table 3: results for pairwise comparisons (statistics)
- Table 4: results for pairwise comparisons (compliance with criteria)
- Table 5: results for BWS comparisons (statistics)
- Table 6: inter-annotator agreement (IAA) scores per subset

4.1 Comparison between GenAI-based and Corpus-based

As appears from Table 3, GenAI-based sentences are more frequently being found suitable than corpus-based sentences, with learners unanimously choosing the artificially generated sentence over the corpus-based one in 265 of the 400 pairwise comparisons (148/200 for the beginner group and 117/200 for the advanced group). In comparison, where the source is corpus-based, this value only amounts to 10/400. The moderate to substantial IAA scores (Table 6) for the corresponding subsets (between 0.62 and 0.72 for beginner and 0.55 and 0.65 for advanced) indicate that these annotations can be considered reliable, especially for the beginner group.

When looking at why corpus-based sentences are found less suitable than their GenAI-based counterparts, Table 4 reveals that – apart from a few cases where they contain the target item in a wrong POS (Example 1) – the corpus examples are less preferred mainly because they are (1) more context-dependent (Example 2, with *la otra* [‘the other’] being dependent on the preceding context) and (2) too difficult (e.g., *rugen* and *se abalanzan* in Example 3). In other words, the selection algorithm based on the SelEjemCor framework sometimes fails to meet the main criterion of *context independence* and the specific criterion of *difficult vocabulary* (see Appendix A). Especially the context dependence of the corpus-based sentences (in 120 of the 400 sentences, i.e. 30%) can be considered an indication that selecting suitable examples from corpora at sentence level is a challenging task. Working at paragraph level might reduce this risk at context dependence (as paragraphs should constitute a more coherent unit of text), but will at the same time also increase the cognitive load and response time of the learning materials based on the examples.

1. Un parlamentario del **tripartito** puso como ejemplo de “buen funcionamiento” y “discreción” la comisión de investigación foral sobre el fraude de la Hacienda de Irún. (‘An MP of the tripartite gave as an example of “good functioning” and “discretion” the foral commission of enquiry into the fraud of the Irún Treasury.’) – Example taken from subset 6 for the adjective *tripartito*

2. Mercedes Alaya instruye ahora además la otra gran **macrocausa** andaluza: el fraude en los cursos de formación. (‘Mercedes Alaya is now also investigating the other big Andalusian mega lawsuit: the fraud in the training courses.’) – Example taken from subset 7 for the noun *macrocausa*

3. En invierno rugen los torrentes que se abalanzan montaña abajo, y el aire fresco agita las **ramas** de los robles. (‘In winter the torrents roar and rush down the mountain, and the fresh air stirs the branches of the oak trees.’) – Example taken from subset 1 for the noun *rama*

In the subsets with BWS evaluations (Table 5), we observe a similar trend: corpus-based examples are more frequently annotated as “worst” (28/50 times by all participants in the beginner group, 26/50 times in advanced) compared to artificially generated examples (2/50 in total for all GenAI prompt types in both beginner and advanced groups). Yet, even though the GenAI approach outperforms the corpus-driven approach by a large margin, Table 4 highlights that there is a non-negligible number of cases where the artificially generated sentences contain the target item in a wrong POS (3 instances in the beginner group, 7 in the advanced group; Example 4), consist of a definition (12 instances in the advanced

Subset	**GenAI-based | Corpus-based**			
	NOUN (/ 30)	VERB (/ 10)	ADJ (/ 10)	Total (/ 50)
1	23 | 0	7 | 0	8 | 1	38 | 1
2	25 | 0	6 | 1	8 | 0	39 | 1
3	24 | 0	6 | 0	6 | 0	36 | 0
4	22 | 1	7 | 1	6 | 0	35 | 2
Total	94 | 1	26 | 2	28 | 1	148 | 4
6	16 | 3	6 | 1	6 | 0	28 | 4
7	19 | 0	6 | 0	3 | 1	28 | 1
8	19 | 0	8 | 0	6 | 0	33 | 0
9	15 | 1	7 | 0	6 | 0	28 | 1
Total	69 | 4	27 | 1	21 | 1	117 | 6

Table 3: Statistics on example sentences annotated as "best" by all participants (N = 3 for subsets 1-4 and N = 4 for subsets 6-9) in pairwise comparison format. Results for the artificially generated sentences appear before the vertical line, results for corpus-based appear after.

	GenAI-based | Corpus-based			
	ZS-G	ZS-G+A	ZS-G+A+C	FEWSH
		Beginner		
Definition	0 | 0	0 | 0	0 | 0	0 | 0
Context-dependent	0 | 16	0 | 13	0 | 15	0 | 17
Too difficult	1 | 11	1 | 12	0 | 11	0 | 9
No target item	0 | 0	1 | 0	0 | 0	0 | 0
Wrong POS	1 | 1	1 | 1	0 | 0	1 | 0
Incomplete	0 | 2	0 | 2	0 | 0	0 | 1
		Advanced		
Definition	3 | 0	1 | 1	5 | 0	3 | 0
Context-dependent	1 | 14	0 | 16	1 | 15	1 | 14
Too difficult	3 | 23	0 | 21	0 | 19	2 | 23
No target item	0 | 0	0 | 0	0 | 0	0 | 0
Wrong POS	2 | 3	2 | 1	3 | 1	0 | 0
Incomplete	0 | 0	0 | 0	0 | 0	0 | 0

Table 4: Details on sentences that did not meet the suitability criteria defined in the annotation instructions, for the pairwise comparison subsets (see also Section 3.4). The number of sentences for GenAI-based appear before the vertical line, the number for corpus-based after the vertical line (on a total of 50, i.e. the number of sentences in a subset). "ZS-G" stands for the ZS-GEN prompt type, "ZS-G+A" for ZS-GEN+AUD, "ZS-G+A+C" for ZS-GEN+AUD+CRIT, and "FEWSH" for FEWSHOT.

		Full agreement | $\geq$ 1 agreement				
		CORP	ZS-G	ZS-G+A	ZS-G+A+C	FEWSH
5_{best}	NOUN	0 | 2	0 | 19	4 | 16	0 | 16	0 | 13
	VERB	0 | 1	0 | 3	2 | 6	1 | 3	0 | 5
	ADJ	0 | 0	0 | 1	0 | 2	2 | 5	4 | 6
	Total	0 | 3	0 | 23	6 | 24	3 | 24	4 | 24
5_{worst}	NOUN	17 | 29	0 | 1	0 | 2	0 | 4	1 | 7
	VERB	7 | 9	0 | 2	0 | 1	0 | 0	0 | 1
	ADJ	4 | 9	0 | 2	0 | 3	0 | 2	1 | 2
	Total	28 | 47	0 | 5	0 | 6	0 | 6	2 | 10
10_{best}	NOUN	0 | 3	0 | 20	0 | 16	1 | 20	2 | 18
	VERB	1 | 1	1 | 7	0 | 5	1 | 4	0 | 4
	ADJ	0 | 1	1 | 5	0 | 5	1 | 7	0 | 6
	Total	1 | 4	2 | 32	0 | 26	3 | 31	2 | 28
10_{worst}	NOUN	17 | 27	0 | 3	1 | 4	0 | 1	0 | 3
	VERB	3 | 8	0 | 1	0 | 3	0 | 1	1 | 2
	ADJ	6 | 10	0 | 2	0 | 2	0 | 0	0 | 3
	Total	26 | 45	0 | 6	1 | 9	0 | 2	1 | 8

Table 5: Statistics on example sentences annotated as "best" and "worst" in subsets 5 (beginner target group) and 10 (advanced). "CORP" stands for corpus-based. The value before the vertical line refers to the sentences for which all of the participants (N = 3 for subset 5 and N = 4 for subset 10) agreed, the value after the vertical line reports the number of sentences for which at least one of the participants chose the sentence. The values in the "Total" rows are on a total of 50 (i.e. the number of sentences in a subset). "ZS-G" stands for the ZS-GEN prompt type, "ZS-G+A" for ZS-GEN+AUD, "ZS-G+A+C" for ZS-GEN+AUD+CRIT, and "FEWSH" for FEWSHOT.

Subset	**IAA (α)**	**ZS-G**	**ZS-G+A**	**ZS-G+A+C**	**FEWSH**	**ALL**
1	0.7	✓				
2	0.72		✓			
3	0.66			✓		
4	0.62				✓	
5_{best}	0.29					✓
5_{worst}	0.61					✓
Avg	0.6	✓	✓	✓	✓	✓
6	0.6	✓				
7	0.58		✓			
8	0.55			✓		
9	0.65				✓	
10_{best}	0.22					✓
10_{worst}	0.71					✓
Avg	0.55	✓	✓	✓	✓	✓

Table 6: IAA scores – as computed by Krippendorff's alpha (α) – for the annotation task in which L2 learners compare corpus-based sentences to artificially generated ones. "ALL" refers to subsets for which an example sentence based on each of the four different input prompts is generated. "Avg" rows report the average IAA value per target group. "ZS-G" stands for the ZS-GEN prompt type, "ZS-G+A" for ZS-GEN+AUD, "ZS-G+A+C" for ZS-GEN+AUD+CRIT, and "FEWSH" for FEWSHOT.

group; Example 5), or are found to be too difficult (2 instances in the beginner group, 5 in the advanced group; Example 6, with the word *desencadenó* being considered difficult by some of the advanced learners). This finding is also backed by the BWS evaluation results in Table 5, which show that there are 27/50 (beginner) and 25/50 (advanced) artificially generated examples annotated as “worse” by at least one of the learners (“$\geq$ 1 agreement”) in total across the four prompt types.

4. **Mañana** vamos a visitar el museo de arte moderno en el centro de la ciudad. (‘Tomorrow we are going to visit the museum for modern art in the city centre.’) – Example taken from subset 2 for the noun *mañana*

5. El **blanqueo** de dinero es un delito grave que involucra la transformación de dinero de origen ilícito en apariencia lícita. (‘Money laundering is a serious crime involving the conversion of money of an illegal nature into a lawful form.’) – Example taken from subset 6 for the noun *blanqueo*

6. La **destitución** del director desencadenó una crisis en la empresa que aún no se ha resuelto. (‘The dismissal of the director triggered a crisis in the company that has not yet been resolved.’) – Example taken from subset 9 for the noun *destitución*

4.2 Comparison between Different Prompt Types

When comparing the full agreement results for the different GenAI prompts in Table 3, there is no noticeable difference (total scores range between 35/50 and 39/50 for the beginner group and between 28/50 and 33/50 for the advanced group). The only values which are slightly out of the ordinary are those for the adjectives in the advanced group: for ZS-GEN, ZS-GEN+AUD+CRIT, and FEWSHOT 6/10 sentences are annotated as “best” by all of the learners, while for the ZS-GEN+AUD prompt type this value only amounts to 3/10. Yet, this evidence is not substantial enough from which to draw conclusions, particularly because ZS-GEN+AUD obtains the top value (8/10) in the corresponding subset for the beginner group (subset 2, ADJ).

The results of the BWS evaluations (Table 5), however, paint a somewhat different picture. For the beginner group, the full agreement scores show that specifying the target audience (ZS-G+A, 6/50 chosen as “best”) and the criteria (ZS-G+A+C, 3/50) has an added value compared to the broad context description (ZS-G, 0/50), just as providing the GenAI model with a few examples (FEWSH, 4/50). Nevertheless, when looking at the “$\geq$ 1 agreement” results, this difference disappears: 23/50 for ZS-GEN and 24/50 for the other three prompt types. Moreover, for the advanced group the ZS-GEN prompt type actually comes out as the arguably second-best prompt type with 2/50 full agreement and 32/50 $\geq$ 1 agreement (compared to 0 and 26/50 for ZS-GEN+AUD, 3 and 31/50 for ZS-GEN+AUD+CRIT, and 2 and 28/50 for FEWSHOT). In other words, even though the BWS evaluations reveal somewhat more outspoken differences, these differences do not follow any clear pattern. This observation is also corroborated by the IAA scores, which are fairly low for the “best” annotations in subset 5 (α = 0.29; beginner group) and 10 (α = 0.22; advanced group).

5 Discussion

Regarding RQ1 (corpus-based versus GenAI as sentence source), the experiment has shown that, overall, L2 Spanish learners find artificially generated example sentences considerably more suitable than corpus-based sentences. The evaluation by the learners revealed that 30% of the corpus sentences were not fully comprehensible without further context. Put otherwise, GenAI methods seem most sensible to use for examples at sentence level, while corpus-based methods might be more suitable to retrieve items in a broader context, for example at paragraph level. However, the results also showed that in a number of cases the L2 learners did prefer the corpus-based example at sentence level, implying that exclusive reliance on GenAI to create sentence-level example sentences is not to be recommended. Moreover, even though the large language models used to generate the artificial examples are trained on large corpora, it is highly questionable if these sentences can be said to represent an authentic expression of language. Therefore, a third method which combines the best of both worlds might be worth considering: starting from a corpus-based example and using a GenAI model to rewrite it.

As for RQ2 (comparison between GenAI prompt types), the results were inconclusive:

adding a higher degree of specificity (by describing the target audience and the criteria the sentence should meet) did not result in any observable improvement compared to using a zero-shot prompt that only sketched the broad context. Opting for a few-shot prompt (i.e. providing a few examples the model can learn from) instead of a zero-shot prompt did not have any noticeable impact on the results either.

A first limitation of the study is that both the dataset size and the number of L2 learners evaluating the example sentences should be increased to arrive at more substantiated conclusions. Furthermore, even though the four different prompts provided considerable variation, more extensive prompt engineering could constitute a valuable avenue for further research, as would the comparison between different large language models for generating the artificial examples. Especially the choice between open-source (e.g., Meta's Llama models) and proprietary/closed-source models (e.g., OpenAI's GPT models) will become one of the most crucial methodological decisions, with the possibility to have a "peek under the hood" being weighed against performance levels and ease of use.

A third potential limitation is that – in the current setup – the target words may appear in a different linguistic construction (e.g., as a part of a collocation/colligation or not), meaning (e.g., literal versus metaphorical sense), or syntactic role (e.g., subject versus object position). It might be argued that differences in these aspects should be limited as much as possible, as they could have an impact on how easy or difficult it is for learners to understand the example sentences. Finally, the role of the texts from which the corpus-based sentences are chosen should also be analysed in further detail, for example by studying if compiling a specific corpus consisting exclusively of texts that have been written for users with a lower proficiency (e.g., from newspapers for children or adolescents) has a positive impact on the corpus-based scores for the beginner group.

6 Conclusion and Future Work

In this paper we compared corpus-based sentences to artificially generated sentences in terms of pedagogical suitability. We constructed a dataset containing 500 target items (250 vocabulary items to be taught to beginner learners and 250 to advanced learners), for which we selected corpus examples according to a dedicated selection algorithm based on the SelEjemCor framework (Appendix A) and generated artificial examples by querying the GPT-3.5 Turbo large language model. The comparative evaluation of the sentences was performed by means of an experiment with seven students of L2 Spanish. The results of the experiment can be summarised into three main takeaways:

1. L2 learners find GenAI-based sentences considerably more suitable than corpus-based sentences. Of the 400 pairwise comparisons between corpus-based and GenAI-based sentences, 265 artificially generated examples were found suitable by all learners, compared to only 10 corpus-based examples.

2. Despite their excellent performance, the use of GenAI models has also shown to yield a number of unsuitable example sentences (with the target word in a wrong POS, the sentence being a definition instead of a usage example, or the sentence containing words that are too difficult).

3. A general zero-shot prompt describing the broad context of the task (i.e. the creation of example sentences for language learning purposes) provides enough information to create suitable example sentences. More specific prompts (describing the target audience and the criteria the sentence should meet) do not lead to better results, nor does formulating the prompt in a few-shot format (i.e. containing a few examples the model can learn from).

In potential follow-up experiments, the limitations discussed in Section 5 should be addressed, starting with increasing the number of target items and participants, evaluating the impact of using different corpora, and applying more extensive prompt engineering based on techniques for educational purposes in general (Cain, 2024) and for L2 learning purposes in particular (Isemonger, 2023). To convert the experimental design adopted in the current study into a more "controlled environment", testing different GenAI models with the same prompts or using designated platforms such as LMStudio are options worth considering. Additionally, fine-tuning the annotation instructions (e.g., by adding an explicit evaluation of the grammatical soundness and syntactic properties of the sentence) would allow us to gain more in-depth insights into the exact reasons why one example sentence is preferred over another.

Furthermore, as hinted at in the discussion (Section 5) as well, developing a new method that combines a corpus-based and GenAI-based approach constitutes another important topic for future research. In such a "hybrid" method, authentic corpus-based examples can be taken as the starting point and GenAI models can be used as the means to rewrite the examples in order to make them meet the required criteria, especially regarding context independence and difficulty. Different types of rewriting prompts could be compared, from zero shot over few shot to retrieval-augmented generation (in which we let the model "look for" the most relevant information in large set of corpus examples and then prompt it to generate new examples based on this information). Yet, our (preliminary) finding that the corpus-based method (yielding *authentic* example sentences) is being outperformed by the GenAI-based one (yielding *artificial* examples) can also be considered a reason to bring that other source of non-authentic examples, the invented example (IE; Section 2.1), back into the equation. Conducting an experiment in which IEs are compared to artificially generated sentences could shed renewed light on the role IEs can play in an L2 setting.

7 Acknowledgements

This research has been carried out as part of a PhD fellowship on the IVESS project (file number 11D3921N), funded by the Research Foundation - Flanders (FWO). Additionally, we want to express our sincere gratitude to the reviewers for their valuable feedback and suggestions.

References

Beryl T. S. Atkins and Michael Rundell. 2008. *The Oxford guide to practical lexicography*, 1 edition. Oxford linguistics. Oxford University Press, Oxford.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

William Cain. 2024. Prompting Change: Exploring Prompt Engineering in Large Language Model AI and Its Potential to Transform Education. *TechTrends*, 68(1):47–57.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. On the application of Large Language Models for language teaching and assessment technology.

G. Cook. 2001. 'The philosopher pulled the lower jaw of the hen'. Ludicrous invented sentences in language teaching. *Applied Linguistics*, 22(3):366–387.

Peter Crosthwaite and Vit Baisa. 2023. Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3):100066.

Mark Davies and Kathy Hayward Davies. 2018. *A frequency dictionary of Spanish: Core vocabulary for learners*, 2 edition. Routledge frequency dictionaries. Routledge, London ; New York.

Nick C. Ellis. 2006. Language Acquisition as Rational Contingency Learning. *Applied Linguistics*, 27(1):1–24.

J.R. Firth. 1968. *Selected Papers of J.R. Firth*. Longman, London; Harlow.

A. Frankenberg-Garcia. 2012. Learners' Use of Corpus Examples. *International Journal of Lexicography*, 25(3):273–296.

Ana Frankenberg-Garcia. 2014. The use of corpus examples for language comprehension and production. *ReCALL*, 26(2):128–146.

Ana Frankenberg-Garcia, Geraint Paul Rees, and Robert Lew. 2021. Slipping Through the Cracks in e-Lexicography. *International Journal of Lexicography*, 34(2):206–234.

Costas Gabrielatos. 2018. Keyness analysis: Nature, metrics and techniques. In C. Taylor and A. Marchi, editors, *Corpus Approaches To Discourse: A critical review*, pages 225–258. Routledge, Oxford.

Patrick Goethals. 2018. Customizing vocabulary learning for advanced learners of Spanish. In *Technological innovation for specialized linguistic domains : languages for digital lives and cultures, proceedings of TISLID'18*, pages 229–240, Gent, Belgium. Éditions Universitaires Européennes.

Maldonado González, editor. 2012. *Diccionario Clave: diccionario de uso del español actual*, 9 edition. SM, Boadilla del Monte (Madrid).

Stefan Th. Gries. 2013. 50-something years of work on collocations: What is or should be next .... *International Journal of Corpus Linguistics*, 18(1):137–166.

Tanja Heck and Detmar Meurers. 2022. Generating and authoring high-variability exercises from authentic texts. In *Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)*, pages 61–71.

Ian Isemonger. 2023. Generative Language Models in Education: Foreign Language Learning and the Teacher as Prompt Engineer. *TEFL Praxis Journal*, 2:3–17.

Jelena Kallas, Adam Kilgarriff, Kristina Koppel, Elgar Kudritski, Margit Langemets, Jan Michelfeit, Maria Tuulik, and Ülle Viks. 2015. Automatic generation of the Estonian Collocations Dictionary database. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, pages 11–13.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1, pages 425–432. Universitat Pompeu Fabra Barcelona.

Lucas Kohnke, Benjamin Luke Moorhouse, and Di Zou. 2023. ChatGPT for Language Teaching and Learning. *RELC Journal*, 54(2):537–550.

Iztok Kosem, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit, and Carole Tiberius. 2019. Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2):119–137.

Batia Laufer. 1992. Corpus-based versus lexicographer examples in comprehension and production of new words. In *Proceedings of the Fifth Euralex International Congress*, pages 4–9. University of Tampere.

Batia Laufer and Karen Shmueli. 1997. Memorizing New Words: Does Teaching Have Anything To Do With It? *RELC Journal*, 28(1):89–108.

I.S.P. Nation. 2022. *Learning Vocabulary in Another Language*, 3 edition. Cambridge University Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Ildikó Pilán, Elena Volodina, and Lars Borin. 2016. Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Revue Traitement Automatique Des Langues*, 57(3):67–91.

Punjaporn Pojanapunya and Richard Watson Todd. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1):133–167.

John McHardy Sinclair. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. Collins ELT, London.

Simon Smith, P.V.S. Avinesh, and Adam Kilgarriff. 2010. Gap-fill tests for language learners: Corpus-driven item generation. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, pages 1–6. Macmillan Publishers India.

Anatol Stefanowitsch. 2020. *Corpus linguistics: A guide to the methodology*. Number 7 in Textbooks in language sciences. Language Science Press, Berlin.

Appendices

Appendix A. SelEjemCor framework

The criteria included in the SelEjemCor are presented in Table 7. The values set for obtaining the example sentences in the experiment are included in the "$\mathrm{V}_{\mathrm{set}}$ BEG" and "$\mathrm{V}_{\mathrm{set}}$ ADV" columns.

Appendix B. Prompt types

The different prompt types used in the experiment are presented in Table 8.

Nr	Criterion	V_{set} BEG	V_{set} ADV
/	Proficiency level target audience.	B1	C1
/	Number of years experience target audience.	1	3
1	Boolean value indicating if search term has to occur in sentence.	*True*	*True*
2	Numerical value indicating maximum number of times search term can occur in sentence.	1	1
3	Numerical value between 0 and 1 indicating at which position search term has to occur.	*None*	*None*
4	Boolean value indicating if sentence has to contain dependency root.	*True*	*True*
5	Boolean value indicating if sentence has to contain subject or finite verb.	*True*	*True*
6	Boolean value indicating if sentence has to contain explicit subject.	*True*	*True*
7	Boolean value indicating if sentence has to start with capital letter and end with punctuation mark.	*True*	*True*
8	Numerical value indicating maximum number of tokens which do not occur in SCAP-based lemma lexicon.	0	1
9	Numerical value indicating maximum number of non-alphabetical tokens (e.g., mark-up traces in web materials).	0	0
10	Boolean value indicating that no conjunction or subjunction can appear in sentence-initial position.	*True*	*True*
11	Numerical value indicating maximum number of demonstrative pronouns (e.g., *este*$\vert$*esta*: 'this'; *ese*$\vert$*esa*: 'that').	0	0
12	Numerical value indicating maximum number of words/phrases which occur in precompiled list of anaphoric expressions (e.g., *allí*: 'there'; *aquí*: 'here'; *entonces*: 'then').	0	0
13	Numerical value indicating maximum number of negation adverbials (e.g., *no*: 'no'; *nadie*: 'nobody'; *nada*: 'nothing').	0	0
14	Boolean value indicating that sentence cannot represent direct question.	*True*	*True*
15	Boolean value indicating that sentence cannot represent direct speech (i.e. speaking verb combined with delimiters such as quotation marks).	*True*	*True*
16	Boolean value indicating that sentence cannot represent answer to closed question (i.e. sentence-initial adverb of affirmation or negation followed by delimiter).	*True*	*True*
17	Numerical value indicating maximum number of tokens which occur in precompiled list of modal verbs (when functioning as an auxiliary verb).	1	3
18	Numerical value indicating maximum number of tokens in the sentence (including punctuation).	10-30	10-30
19	Numerical value indicating maximum number of words above the proficiency level of the target audience according to a personalised machine learning classifier.	0	0
20	Numerical value indicating minimum frequency of words in SCAP lemma frequency dictionary (expressed in percentiles).	P90	P75
21	Numerical value indicating maximum number of words not included in SCAP token lexicon.	0	1
22	Boolean value indicating that sentence cannot contain words which occur in precompiled list of potentially sensitive words related to PARSNIP topics.	*True*	*True*
23	Numerical value indicating maximum number of proper names.	2	2
24	Numerical value indicating minimum average normalised Lexicographer's Mutual Information (Bouma, 2009) score for verb-noun pairs (in subject, object, and oblique relation) and all noun-adjective pairs (in attributive or predicative relation) in the sentence. The scores are retrieved from a SCAP-based resource.	*all*	*all*
25	Numerical value indicating minimum average ΔP score (Ellis, 2006; Gries, 2013) for verb-noun pairs (in subject, object, and oblique relation) and all noun-adjective pairs (in attributive or predicative relation) that include the search term. The scores are retrieved from a SCAP-based resource.	*all*	*all*
26	Numerical value indicating minimum average cosine similarity of serch term with head and dependents (both static and contextualised word embeddings).	*all*	*all*
27	Numerical value indicating minimum average *n*-gram frequency of the sentence (excluding *n*-grams with punctuation marks). Frequencies are retrieved from SCAP dictionary containing lemma-based *n*-grams.	*all*	*all*

Table 7: Criterion descriptions and Values set for SelEjemCor criteria. Filters are put in bold, rankers in plain text. "BEG" and "ADV" refer to the beginner and advanced target groups respectively.

Prompt ID	Prompt text
SYS1	You are a teacher of Spanish as a foreign language.
SYS2	You are a teacher of Spanish as a foreign language to a beginner/lower-intermediate group of university students who have been studying Spanish for one year.
SYS3	You are a teacher of Spanish as a foreign language to an upper-intermediate/advanced group of university students who have been studying Spanish for three years.
USR1	Write a sentence between 10 and 30 words in Spanish that presents an authentic usage of the Spanish [POS] '[WORD]', a vocabulary item that has to be learnt by your students. The sentence should not be a definition of the word.
USR2	Write a sentence between 10 and 30 words in Spanish that presents an authentic usage of the Spanish [POS] '[WORD]', a vocabulary item that has to be learnt by your students. The sentence should not be a definition of the word. The sentence should be well-formed and context-independent, it should be tailored to the proficiency level of your students, and it should contain phrases that frequently co-occur with the target item '[WORD]'.
USR3	Write a sentence between 10 and 30 words in Spanish that presents an authentic usage of a Spanish vocabulary item that has to be learnt by your students: word=diseño; part of speech=noun; sentence=Para hacer un buen diseño de un mueble hay que pensar en su utilidad. ### word=comprometer; part of speech=verb; sentence=Sus revelaciones comprometían en el caso de corrupción a otras dos organizaciones. ### word=dramático; part of speech=adjective; sentence=Toda la prensa se hace eco del dramático caso de la niña desaparecida. ### word=[WORD]; part of speech=[POS]; sentence=

Prompt type	System role	User role	Subset
		Beginner	
ZS-GEN	SYS1	USR1	1
ZS-GEN+AUD	SYS2	USR1	2
ZS-GEN+AUD+CRIT	SYS2	USR2	3
FEWSHOT	SYS2	USR3	4
		Advanced	
ZS-GEN	SYS1	USR1	6
ZS-GEN+AUD	SYS3	USR1	7
ZS-GEN+AUD+CRIT	SYS3	USR2	8
FEWSHOT	SYS3	USR3	9

Table 8: Detailed overview of prompt types used to generate artificial example sentences.