COCOTEROS: A Spanish Corpus with Contextual Knowledge for Natural Language Generation

María Miró Maestre, Iván Martínez-Murillo, Elena Lloret, Paloma Moreda and Armando Suárez Cueto

Dept. of Software and Computing Systems, University of Alicante, Apdo. de Correos 99, E-03080, Alicante, Spain

Abstract

Contextual information is one of the key elements when automatically generating language with a more semantic-pragmatic perspective. To contribute to the study of this linguistic aspect, we present COCOTEROS, a COrpus of COnTextual TExt geneRatiOn in Spanish. COCOTEROS is available at https://huggingface.co/datasets/gplsi/cocoteros. The corpus is composed of sentences and automatically generated context pairs. For creating it, a semi-automatic weakly supervised methodology is implemented. Taking as a reference the Spanish section of the Tatoeba dataset, we filtered the sentences according to our research purpose. Then, we determined several linguistic parameters that the generated contexts need to fulfil considering their reference sentence. Finally, contexts were automatically generated using prompt engineering with Google's large language model Bard. Furthermore, we performed two types of evaluation to check both the linguistic quality and the presence of gender bias in the corpus: the former by manually measuring the magnitude estimation metric and the latter thanks to the GenBit automatic metric. The results show that COCOTEROS is an appropriate language resource to approach Natural Language Generation tasks from a semantic-pragmatic perspective for Spanish. For instance, the NLG task of concept-to-text generation could benefit from contextual information by generating sentences according to the information provided in the context and a set of given concepts. Additionally, regarding the task of question-answering, the inclusion of linguistic context can enhance the generation of more appropriate answers by serving as a guide on what information to include in the automatically generated answer.

Keywords

corpus, contextual information, natural language generation, Spanish, human evaluation, large language models

1. Introduction

Natural Language Generation (NLG) systems are steadily improving their performance in a wide range of tasks where the information to be generated is delimited according to the objective of the task, e.g., text summarisation, machine translation or question answering (QA). One of the most important issues those systems have to deal with is the lack of sufficient contextual knowledge, as it prevents NLG models from better adapting the generated text to the communicative situation of each task. That derives in crucial problems such as the hallucination issue and lack of commonsense in the produced text [1]. In fact, one of the current concerns within the NLG discipline [2] is the need to address tasks from a more 'semantic-pragmatic perspective' to solve these contextual inference difficulties that affect the output of the systems at issue. To address the lack of studies that bear in mind these linguistic levels of analysis, NLG is starting to put linguistic context in the research spotlight, given its importance for appropriately understanding human utterances. Indeed, Newman et al. [3] already defended the consideration of context not only to create text automatically but also to assess the suitability of the generated text. This statement comes from the idea that communication-based features help to evaluate the performance of any model that imitates human language. Language itself is used to communicate ideas always expressed within a given communicative context, and it is such context what directly affects the structure of the utterance we want to say.

Parallel to this, making NLG systems aware of contextual knowledge involves the creation of new resources, such as datasets, corpora, knowledge bases, etc., to train models in several languages, especially for those different from English or low-resourced ones. In the case of Spanish, we observed that most of the recently published corpora hardly address pragmatic-related issues with a contextual perspective, but rather focus on concrete pragmatic aspects such as metaphors to tackle identification tasks. Furthermore, the high performance of Large Language Models (LLMs) recently witnessed within the field of Natural Language Processing (NLP) has allowed researchers to use NLP tools to automatise data collection and corpus creation tasks, therefore reducing the time spent in collecting sufficient data for research purposes

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.

maria.miro@ua.es (M. M. Maestre); ivan.martinezmurillo@ua.es (I. Martínez-Murillo); elena.lloret@ua.es (E. Lloret); moreda@ua.es (P. Moreda); armando.suarez@ua.es (A. S. Cueto)

0000-0001-7996-4440 (M. M. Maestre); 0009-0007-5684-0083 (I. Martínez-Murillo); 0000-0002-2926-294X (E. Lloret); 0000-0002-7193-1561 (P. Moreda); 0000-0002-8590-3000 (A. S. Cueto)

[4, 5].

To bridge the gap of NLG systems that handle more semantic-pragmatic features of language, specifically contextual knowledge, we present COCOTEROS, a COrpus of COntextual TExt geneRatiOn in Spanish. This corpus comprises 4,845 sentences extracted from the existing Tatoeba dataset, together with 4,845 context sentences automatically generated with Bard[1] language model and manually revised. Given the difficulties inherent to prompt engineering when using LLMs-based chatbots, several linguistic parameters were determined to ensure the quality of the automatically generated outputs with Bard, including semantic similarity, length of the generated text, and forbidden keywords, among others. Moreover, we performed a human evaluation experiment based on the magnitude estimation metric with three linguistics specialists to measure the contextual appropriateness of the resulting contexts. In parallel, we measured gender bias with the GenBit tool [6] to verify that our corpus would be useful for NLG tasks without adding gender biases to models trained in further experiments.

In sum, the **main contributions** of this paper are:

- Expansion of a subset of Tatoeba's corpus with contextual information.
- Proposal of a weakly supervised methodology for building a corpus using prompt-engineering.
- Creation of COCOTEROS, a novel Spanish corpus for commonsense NLG that includes contextual information.
- Corpus validation through human assessment with the magnitude estimation method.
- Corpus evaluation of gender bias with GenBit automatic metric.

We believe this corpus will provide the research community with a valuable resource in Spanish to test the performance of NLG systems in different tasks by considering semantic-pragmatic aspects of communication as contextual appropriateness. Some of the NLG tasks that could use COCOTEROS could be those related to concept-to-text generation, where sets of words are provided and the model has to generate a text given those concepts. These words can have multiple semantic meanings depending on their context. Having a prefixed context within which the sentence has to be generated could help to a more precise sentence. Moreover, COCOTEROS could also be used to train NLG models to automatically generate sentences in accordance with a given context as input information, therefore improving the model's awareness of the different communicative situations it is trained with. As for NLP, some tasks that have already exploited the role of contextual knowledge for improving their classification systems are Named Entity Recognition (NER) [7], word recognition and lexical processing to boost semantic disambiguation [8], language learning [9] or even healthcare studies devoted to diseases or syndromes which critically affect language [10].

[1]Since 8th February 2024 Bard is known as Gemini.

2. Related Work

In view of the multidisciplinary nature of the task, the following theoretical background is on one side focused on the linguistic notion of context and its approach to NLG research (subsection 2.1). On the other side, subsection 2.2 includes prior NLG research focused on the creation of linguistic resources to address contextual-related tasks.

2.1. Linguistic Context in NLG Research

Messages need their surrounding communicative context in order to be completely understood [11]. This claim is well accepted within the NLP discipline, as many tasks try to solve context-related linguistic issues to improve NLP systems performance, i.e., coreference resolution [12], information retrieval [13], word sense disambiguation [14] or question answering [15]. Context, therefore, becomes a pragmatic element of great interest when processing language automatically. Similarly, when focusing on language generation, there are concrete applications such as dialogue systems where context is usually predetermined so researchers can study the linguistic features surrounding such communicative context [16, 17].

When addressing the task of contextual appropriateness (i.e., how appropriate is a context given a linguistic setting), several conceptions of context may come to mind, as linguistic theories tend to diverge on the definition of context given the wide range of perspectives from which context can be approached [18]. For the sake of the present research, we focus on the linguistic context of a given message, which can be defined as 'any contiguous span of text within a document' [19] or as 'the set of utterances that precedes the current one' [20]. These definitions align with the linguistic dimension of context known as 'intratextual context' (or 'co-text'), which studies the relation of a piece of text to its surrounding text [18].

2.2. Contextual Corpora for NLG

The creation of linguistic resources directly oriented to analyse more complex linguistic phenomena such as context provides an added asset value to the research community, as there are not as many resources available as for other far-reaching linguistic levels of analysis as syntax or grammar. To motivate the study of this pragmatic element, several resources to analyse context from different perspectives have already been made

available. Castilho et al. [21] created an English corpus annotated with context-aware issues for the task of Machine Translation into Brazilian Portuguese. Regarding dialogue tasks, Udagawa and Aizawa [22] addressed the common grounding problem by collecting a dialogue dataset with continuous and partially-observable context. As for controllable text generation, Lin et al. [23] created the CommonGen task and dataset to test to which extent a generation system can generate text with commonsense reasoning in English. To this end, the task is to generate a coherent sentence that includes several common concepts previously shown to the system. Derived from this work, Carlsson et al. [24] generated the C2Gen dataset of context sentences in English from which they extracted several keywords that had to be included in an automatically generated text. Finally, a recent English corpus worth mentioning is databricks-dolly-15k, a human-generated instruction corpus created to train Dolly LLM [25]. This dataset was applied to different contextual tasks such as summarisation of Wikipedia articles or closed QA, where a question and a reference passage are input to the system to get factually correct responses.

Focusing on Spanish resources, Sanchez-Bayona and Agerri [26] generated a corpus of Spanish metaphors, which depend directly on the contextual meaning to be clearly identified by an automatic system. As for Natural Language Inference (NLI), Kovatchev and Taulé [27] compiled the INFERES corpus to check the performance of machine learning systems on negation-based adversarial examples by using context paragraphs from topics extracted from the Spanish Wikipedia.

After a thorough review of the current corpora that address contextual NLG tasks in Spanish, we can say that, to the best of our knowledge, there is no corpus focused on the contextual information generation task in Spanish. Consequently, for this research we base on the previous works by Lin et al. [23] and Carlsson et al. [24] to address the task of contextual information generation in Spanish.

3. Corpus Creation

The following subsections include the methodology steps to create COCOTEROS: i) we explain the reference sentences dataset collection process and how we filtered them (subsection 3.1); ii) we move on to determine the linguistic constraints that will comprise the prompt to generate automatic contexts (subsection 3.2); iii) we describe the context generation task (subsection 3.3); and iv) we include a manual post-edition to curate the results generated by the LLM (subsection 3.4). Figure 1 shows a visual pipeline of the methodology used for creating COCOTEROS.

3.1. Data Collection and Filtering

For the present study, we wanted to gather simple Spanish sentences with enough semantic content to automatically generate contexts linked to the situation stated in the reference sentence. We prioritised sentences with not too much linguistic information so the context does not add extra information besides the purpose of the task, being not too distant from the original sentence situation. To this end, we chose the Spanish section of sentences written on the website Tatoeba[2] as the original dataset from which we would select the sentences to generate the contexts. We first considered using other already published corpora such as CommonGen [23] or C2Gen [24] as original datasets because they also focused on the task of NLG with contextual information. However, for using these corpora we would have had to translate the original datasets into Spanish, which would imply choosing an appropriate automatic translation tool or manually translating the datasets for adapting the task into Spanish. Also, a further proofreading step would have been necessary to check the accuracy of the translations into Spanish, so we preferred to benefit from an already-existing Spanish dataset that could help us generate our context corpus.

Tatoeba's original dataset includes around 393,000 Spanish sentences either translated from other languages or directly written in Spanish. The dataset includes sentences with a range of 1 up to 44 words per sentence, so we first filtered them by selecting only those sentences conformed by either 8 or 9 words, collecting a total of 60,170 Spanish sentences. We chose this section from the dataset after a previous preprocessing of an excerpt of the dataset with Spacy tokenizer[3]. In this preliminary preprocessing, we noticed that the more words the sentence comprised, the more risk we had of including too much semantic information in the sentence. This could entail the generation of contexts not linked to the original situation stated in the reference sentence. Similarly, we rejected those sentences made up of 7 words or less, as many of their keywords lacked enough linguistic information (verbs, nouns, etc.) to generate a context that could be in line with the situation stated in the reference sentence.

3.2. Linguistic Constraints

LLMs can be useful for supporting the automatic creation of corpora to study specific linguistic phenomena that would become very costly tasks if compiled manually. Nevertheless, generating a corpus with LLMs from scratch also entails several risks regarding linguistic ap-

[2]This dataset was released under a CC-BY License and can be found at https://tatoeba.org/es.

[3]https://spacy.io/api/tokenizer

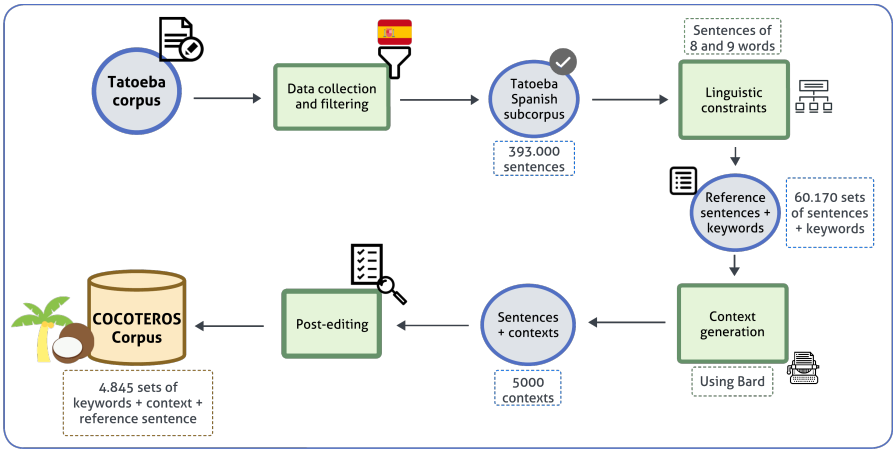

Figure 1: Proposed methodology pipeline for corpus creation.

propriateness that could worsen the quality of the corpus, as it happens with hallucination issues or lack of commonsense.

Therefore, with the aim of automatically creating linguistic contexts referred to a given sentence, and to better control the output of our chosen LLM (further explained in Section 3.3), we determined several linguistic parameters to include in the prompt:

- **Definition of context**: Following previous studies focused on context as described in Section 2.1, we started our prompt with a simple and straightforward definition of what we consider a linguistic context so the model could first get the idea of the task to accomplish.
- **Reference sentence or synonyms**: On the first attempts to find the right prompt to compile the corpus, we observed that, even by including a short definition of linguistic context, the model sometimes generated a context including the reference sentence. Therefore, to better specify the linguistic nature of the context to be generated, we indicated that the reference sentence could not appear in the context nor a sentence with similar semantic meaning.
- **Forbidden keywords**: We extracted three keywords from each reference sentence that could semantically define the sentence meaning. The extraction was automatically performed by means of a random choice where we prioritised the selection of two nouns and one verb, as we consider them some of the main linguistic elements that define the semantic meaning of a sentence. Then, we added those batches of three keywords in the prompt as forbidden words that could not appear in the context to be generated. With this restriction, we wanted to ensure that, even if some of the linguistic structures in the reference sentence were repeated in the generated context, the semantic meaning of the context is related to, but changes somewhat from the reference sentence. This goes in line with the idea that the choice of words influence co-text and meaning potential [28], and we wanted to test up to which point can LLMs generate co-text with the same conceptual background but adding new words that can enlarge the semantic information of the new sentence.
- **Maximum context length**: Inspired by the work presented in Carlsson et al. [24], we decided that an appropriate length for the generated context could be around 45 words. This decision comes also from preliminary prompt tests where we found that, if no length limitation was included, the model tended to delve into the generation process, creating contexts of more than ten lines of text that distanced too much from the original situation stated in the reference sentence.

3.3. Context Generation

Once we filtered the original dataset, the next step was to generate an appropriate context for each of the selected sentences. For this, we benefited from the capabilities of LLMs, and in particular, we used Bard [29], Google's recent LLM. Our decision was motivated by an empirical study we previously conducted in which several LLMs were compared to check how appropriately they fulfilled the task of generating a context resembling a sentence but without repeating or paraphrasing it. The LLMs

compared were LLaMa[4] [30], Vicuna[4] [31], Bard, and ChatGPT[5] We automatically generated contexts for our subset of sentences of 8 or 9 words with Bard, which could generate a context in an average of 5 seconds. Nevertheless, Bard's public version could be prompted only 130 times per day. The generation process was made through a zero-shot prompt that comprised the linguistic restrictions the generated context should include or not, as stated in section 3.2. With this setup, we created an initial version of COCOTEROS corpus with 5,000 contexts.

3.4. Post-editing

Finally, given Bard's predefined chat-like communicative structures, we manually revised and post-edited the resulting contexts by eliminating all the information included in the response which was not the generated context itself (e.g. Bard's output included similar sentences to *"Aquí tienes un contexto relacionado para la frase 'Tengo demasiadas cosas en la cabeza estos días'"* as a preliminary statement for each context[6]). As a remark, there were times when Bard generated several contexts for a single input, giving us the opportunity to choose between them, so we did a manual proofreading process where we checked every possible context to choose those that approximated more to the conception of context we determined for this research task. In line with this, in those cases where we could choose from two options, we selected the context describing a female-subject situation. We made this decision because we detected a somewhat higher proportion of reference sentences addressing male subjects, so the generated context was male-gendered too. Therefore, in those cases where the reference sentence was no gender-specific, we prioritised female contexts to balance gender in COCOTEROS. Further details on how we addressed gender bias in our corpus are shown in subsection 5.2.

In this manual post-editing step we also discarded contexts that were repetitions or paraphrasing of the reference sentence, as well as those that did not include enough semantic information to be considered appropriately generated contexts. Within the rest of contexts we kept in COCOTEROS, there were times where Bard left some of the concepts in the generated text incomplete so the user could complete it according to his/her preferences, as in "*Nos encontramos a [nombre de la persona] sentado en su escritorio*"[7]. Consequently, we had to modify those contexts by completing the missing information with generic concepts or names so we could add the resulting context to the final corpus.

4. COCOTEROS - Corpus of Contextual Text Generation in Spanish

As the first corpus focused on the contextual text generation task for Spanish, COCOTEROScontains a total of 4,845 pairs of reference sentences with their respective generated contexts as illustrated in Figure 2. Moreover, the corpus includes the three keywords extracted from each reference sentence. The final amount of contexts comes from a previous manual post-edition from the original 5,000 contexts generated with Bard. We performed this post-edition because we noticed sexist content in some of the generated contexts, so we decided to discard those cases straightforwardly.

Table 1 shows a statistical summary of COCOTEROS. Apart from the corpus general information, we found it interesting to check the average sentences and words per context because Bard sometimes generated contexts with very different lengths. Even though the prompt included the maximum length that the context could have (45 words), we found cases where the context had only 15 words, whereas other contexts contained more than four sentences, with a total of more than 50 words.

Table 1
COCOTEROS data summary.

Data	Total
Reference sentences	4,845
Keywords	14,535
Generated contexts	4,845
Words in the sentences	40,827
Words in the contexts	119,885
Words in the corpus	175,247
Average no. of sentences per context	2
Average no. of words per context	25

The official version of COCOTEROS corpus is available at https://huggingface.co/datasets/gplsi/cocoteros. With this, we aim to contribute to NLG research with a new language resource for studying contextual information generation in Spanish, as well as for other unexplored NLG tasks that can benefit from our corpus to address further research questions.

[4]The tested version of LLama was llama-2-70b-chat, and Vicuna's version was vicuna-33b. They were tested on https://chat.lmsys.org/

[5]Tested version of ChatGPT was GPT 3.5 on https://chat.openai.com/

[6]Example translated into English for clarity purposes: Here's a related context for the phrase "I have too many things on my mind these days"

[7]Example translated into English for clarity purposes: We found [name of the person] sit on his/her desk

Reference sentence	Keywords	Generated context
Debes dejar de actuar de forma tan tonta. --- *You should stop acting in such a dumb way.*	['dejar', 'actuar', 'forma'] --- *['stop', 'act', 'way']*	Tu comportamiento es inapropiado y está perjudicando tu reputación. Por favor, sé más maduro y responsable. --- *Your behaviour is inappropriate and it's damaging your reputation. Please, be more mature and responsible.*
El vecindario entero se sorprendió por la noticia. --- *The whole neighbourhood was surprised by the news.*	['sorprender', 'vecindario', 'noticia'] --- *['surprise', 'neighbourhood', 'news']*	La noticia del cierre de la escuela fue como un jarro de agua fría para los vecinos. --- *The news of the school's closure came as a complete shock to the neighbours.*
Te prometo que las cosas van a cambiar. --- *I promise you things are going to change.*	['prometer', 'cambiar', 'cosa'] --- *['promise', 'change', 'things']*	Te aseguro que todo va a mejorar. Estoy trabajando duro para que eso suceda. --- *I assure you that everything will get better. I'm working hard to make it happen.*

Figure 2: Excerpt of COCOTEROS corpus. Examples translated into English for clarity purposes.

5. Corpus Evaluation

To ensure that the contexts included in COCOTEROS are appropriate for contextual generation tasks, we evaluated them taking into account different aspects: context appropriateness with the manual magnitude estimation method (subsection 5.1) and gender bias through the automatic GenBit metric (subsection 5.2).

5.1. Context Appropriateness

With the evolution of the latest LLMs, researchers face a need for consistent evaluation metrics that help them evaluate the outputs provided by these models when testing their performance for language generation tasks. To this end, we performed an experiment based on the magnitude estimation method [32] with the help of three linguistics specialists. Magnitude estimation is a method generally used in psychology to check the reaction of different subjects when presented with several stimuli. To measure the different levels of reaction subjects can have, they need to assign a score to a first stimuli (in our case, the generated context) where no ranges or limits are determined. Then, when a second stimuli is presented, they have to compare it with the first stimuli shown, and depending on the intensity of the reaction they have, its score will change based on the previous score they assigned to the first stimuli. In this manner, if subjects' reaction to the second stimuli is twice as much as to the first stimuli, they will have to double the score they assign to the second stimuli. This method has been used positively for evaluating automatically generated text in several NLG tasks [33, 34, 35, 36], as researchers demonstrated that it helps to detect more linguistic nuances as well as more distinctive rankings when comparing the outputs between the annotators in comparison to other more common methods such as Likert scales [33, 35].

Taking this method as a basis, we wanted to measure the appropriateness of the generated context given its reference sentence. For this, we took a representative sample of sentences and contexts from the COCOTEROS corpus through Formula 1, presented in [37] and previously used in [38]:

$$M = \frac{N * K^2 * P * Q}{E^2 * (N-1) + K^2 * P * Q} \quad (1)$$

where N is the population, K the confidence interval, P the probability of success, Q the probability of failure and E the error rate. The population N was 4,845 sentences and their respective contexts, and the values given to the rest of these parameters were taken as presented in [39], so that K=0.95, E=0.05, P=0.5, and Q=0.5. Once the formula was calculated, the resulting number of sentences M for testing contextual appropriateness was rounded to 90 sentences with their respective contexts. This subset of 90 sentences and contexts was selected at random from the final COCOTEROS corpus. With the subset of contexts already determined, we performed the magnitude estimation analysis to validate the generated contexts.

To accomplish this, we explained the methodology to score the subset of 90 generated contexts to the annotators, with the only requirement that the lowest score they could assign could be 1. In this manner, we ensured the subsequent normalisation of the values each of them may assign to each context. As a remark, we noticed that two annotators scored contexts based on a 1 to 100 ranking, even when we highlighted that there were no restrictions in the values they could choose for

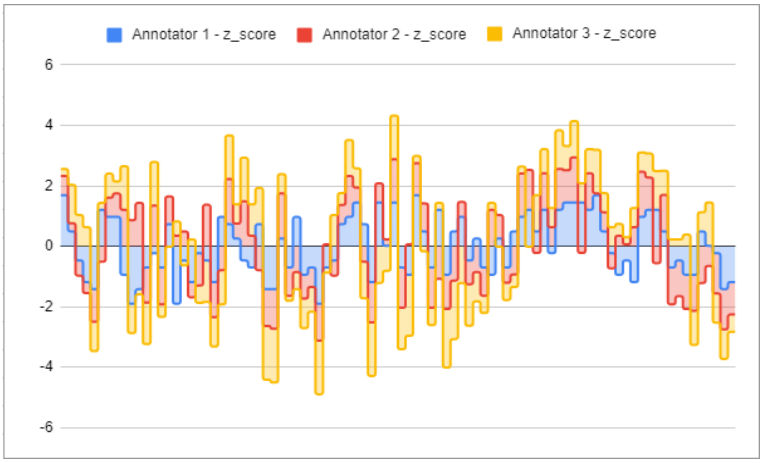

Figure 3: Results of Z-score normalised values for the magnitude estimation evaluation. Values higher than 0 indicate appropriate contexts, whereas negative values show not-suitably generated contexts.

each context. Once we collected all the scores made by the annotators, we normalised the results by means of the z-score normalisation formula (Formula 2) as used in [40]:

$$Z_{ih} = \frac{x_{ih} - \mu_h}{\sigma_h} \tag{2}$$

where Z_{ih} is annotator h's z-score for the context when annotator h gave a magnitude estimation score of x_{ih} to that context. μ_h is the mean and σ_h the standard deviation of the set of magnitude estimation scores for annotator h.

Figure 3 shows the normalised results for the magnitude estimation evaluation. The *0* line serves as the mean from which upper numbers indicate those contexts with higher scores, and the negative numbers show those contexts considered not suitably generated. As can be seen, the three annotators tend to agree on which linguistic contexts have an appropriate contextual relatedness to the reference sentence, even though each of them used a different range of scores within the magnitude estimation experiment. In spite of a few disagreeing cases in the total of 90 contexts, we observe that the annotators agree that more than half of the corpus sample comprises contexts with appropriate contextual relatedness, while the rest could be improved. After evaluating the results with the annotators, we concluded that they tended to highly penalise those contexts that paraphrased the reference sentence, even if after that paraphrasing sentence the context included new excerpts of text that indeed served as an appropriate linguistic context.

5.2. Gender Bias

Several methodological issues come to mind when using a LLM to generate a new language resource for further training LLMs so they can learn how to approach new emerging NLG problems. One of those recently detected issues is the presence of gender bias in the human-compiled corpora that LLMs are trained with. This poses a new problem for the research community, as the incredible performance those LLMs currently show is based on data that reflect and amplify societal biases detected in naturally occurring texts [41]. With an eye to check possible biases in our corpus, we used the Gender Bias Tool (GenBit) [6] to measure the apparent level of gender bias in the 4,845 generated contexts from COCOTEROS. According to its developers, GenBit helps determine if gender is distributed uniformly across data by measuring the strength of association between a pre-defined list of gender definition words and other words in the corpus via co-occurrence statistics. Table 2 shows the obtained results after processing COCOTEROS with the Spanish metric provided in GenBit.

Table 2
GenBit gender bias results in the generated contexts.

Metric	Results
GenBit Score	0.724
Female words	0.335
Male words	0.665

Following the benchmarks as stated in Sengupta et al. [6], the GenBit score from COCOTEROS is 0.724, which indicates a moderate gender bias in our corpus. This key metric comes from a parallel calculation where GenBit calculates the percentage of female or male-gendered

definition words that appear in the corpus, resulting in 0.335 and 0.665, respectively, in COCOTEROS. Considering the results, it seems there is a higher representation of words associated with the male gender rather than with the female. However, these results do not imply that those sentences containing female-gendered words are used in a sexist context but that the appearance of female-gendered terms in the corpus is lower. We want to remark on this because the apparent underrepresentation of female-gendered words could be modified easily by creating parallel contexts to those where there are male-gendered words so that we could balance both gender representation at the same time that we expand our corpus with further examples. Moreover, we have to bear in mind that words in Spanish have a specific genre, whereas English words don't. Consequently, a predominance of male-gendered words does not need to imply that the corpus is gender biased, but that the corpus includes more words linked to that genre, whether those words refer to objects, places or people.
In addition, during our manual post-editing stage of the 4,845 contexts, we found that many of them described communicative situations where the subject is a woman. However, GenBit does not include female or male proper names and gendered adjectives in its Spanish section, so it cannot consider those contexts as gendered-defined, which may also affect the final result of the gender bias metric. Therefore, the results achieved with GenBit score serve as a first attempt to consider possible gender bias in our corpus, but we believe they cannot be conclusive given the different examples of gendered sentences found in our corpus not considered by the metric.

6. Overall Discussion

The results obtained throughout the experimentation process for creating and evaluating COCOTEROS open the door for discussion along several dimensions.

Regarding the magnitude estimation evaluation, this metric helped us to detect further nuances in the scores each annotator assigned to contexts depending on their appropriateness. Those nuances could be future challenges to address to keep on discovering knowledge on how to deal with contextual information in NLG systems. Therefore, these results helped us to determine one of the modifications to apply to COCOTEROS, as in future work we will manually analyse and discard contexts with paraphrasing sentences, so we only leave linguistic contexts that add contextual information to the reference sentence without using synonyms.

Another key aspect of generating new resources is that they must not contain gender biases. An unbiased dataset is an important factor when training a language model, as bias is mostly introduced in the data used in the training phase of the model. As discussed in Section 5.2, remarkable efforts have been made to balance the number of sentences addressing both genders as we are aware of the importance of dealing with gender underrepresentation when creating inclusive language resources that comply with gender balance standards. By doing this, we also want to encourage the rest of the community to take similar steps so that NLP resources and LLMs are trained on trustful resources with no biases. We used GenBit tool for measuring this number, and although the results obtained are the expected, it is true that GenBit does not detect some grammatical categories such as male or female proper names and gendered adjectives, so the results cannot be conclusive.

One problem worth commenting regarding LLMs is hallucination, which occurs when a text is nonsensical or unfaithful to the input source. During post-processing, we detected that some generated context suffered from this (e.g., the reference sentence contained the word "father" while the context was generated with "grandfather"; the generated context was written in the masculine form when the reference sentence was in the feminine form; or the case of fake generated data, such as the winner of Eurovision 2023 which was not Germany). Nevertheless, we did not discard these sentences as our scope was to obtain appropriate contexts. Therefore, future works will focus on detecting and eliminating hallucinations to gather a corpus free of this issue.

Finally, another of the main interests for generating new resources for the NLP community is creating multitask datasets so that linguistic resources become a valuable and reusable tool which can motivate new research. COCOTEROS will contribute to boosting NLP research specifically addressing semantic and pragmatic aspects and for Spanish language. Although it has been originally conceived for NLG, its nature for containing contexts associated with reference sentences could be beneficial for solving other NLP-related issues such as textual entailment, also known as Natural Language Inference (NLI) [42]. This task focuses on the semantic relations that may exist between several pieces of text and how such relations can be characterised and computationally analysed.

7. Conclusions and Future Work

In this paper we have presented COCOTEROS, a Spanish corpus of contextual knowledge for NLG, containing nearly 5,000 sentences with their corresponding contexts. The creation of COCOTEROS comes from the current need in NLP research to address tasks with a more semantic-pragmatic approach, as it occurs with the generation of linguistic context. Also, we wanted to contribute to the research community with a well-defined Spanish

resource to study contextual aspects in NLG, given the lack of enough linguistic resources to study pragmatic aspects of language for languages other than English.

With the aim of verifying the level of linguistic and contextual appropriateness of COCOTEROS, we performed a two-fold evaluation. First of all, we used the magnitude estimation method with the help of three linguistics specialists to measure the linguistic and contextual appropriateness of a representative sample of the generated contexts. Then, we applied the GenBit metric to COCOTEROS to check the level of gender bias our corpus showed. On the one hand, results on the contextual appropriateness evaluation reflect the difficulties when addressing the contextual generation task even for human annotators, as annotators tended to differ on the degree of appropriateness of each context. Nevertheless, the magnitude estimation metric indicates that more than half of the evaluated contexts were scored favourably. On the other hand, the gender bias metric score shows that, with a few modifications, we could reduce the presence of gender bias in the corpus to a large extent. However, the resulting bias score cannot be conclusive as the metric did not consider some of the gender-linguistic features the generated contexts included.

Several research directions are planned for future work. First, we would like to improve our resource, so further experiments will be made to balance gender representation in COCOTEROS, as well as to extend the number of contexts so this Spanish resource may be of help for addressing NLP tasks that need more amounts of data. Finally, we aim to devote a branch of future research to adapting COCOTEROS corpus to the task of intention identification to better understand which reasons make humans have a particular intention when uttering a message based on the context surrounding such intention. At the same time, we would check if LLMs can better detect specific communicative intentions depending on reference sentences and their linguistic context.

Acknowledgments

The research work conducted is part of the R&D projects “CORTEX: Conscious Text Generation” (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by “ERDF A way of making Europe”; “CLEAR.TEXT:Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities” (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and “European Union NextGenerationEU/PRTR”; and the project “NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation” with grant reference (CIPROM/2021/21) funded by the Generalitat Valenciana. Moreover, it has been also partially funded by the Ministry of Economic Affairs and Digital Transformation and “European Union NextGenerationEU/PRTR” through the "ILENIA" project (grant number 2022/TL22/00215337) and "VIVES" subproject (grant number 2022/TL22/00215334).

References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023). URL: https://doi.org/10.1145/3571730. doi:10.1145/3571730.

[2] M. a. T. Yan Li, D. Liu, From semantics to pragmatics: Where IS can lead in natural language processing (NLP) research, European Journal of Information Systems 30 (2021) 569–590. doi:10.1080/0960085X.2020.1816145.

[3] B. Newman, R. Cohn-Gordon, C. Potts, Communication-based evaluation for natural language generation, in: Proceedings of the Society for Computation in Linguistics 2020, Association for Computational Linguistics, New York, New York, 2020, pp. 116–126. URL: https://aclanthology.org/2020.scil-1.16.

[4] J. C. B. Cruz, J. K. Resabal, J. Lin, D. J. Velasco, C. Cheng, Exploiting news article structure for automatic corpus generation of entailment datasets, in: D. N. Pham, T. Theeramunkong, G. Governatori, F. Liu (Eds.), PRICAI 2021: Trends in Artificial Intelligence, Springer International Publishing, Cham, 2021, pp. 86–99.

[5] M. E. Vallecillo-Rodríguez, A. Montejo-Raéz, M. T. Martín-Valdivia, Automatic counter-narrative generation for hate speech in spanish, Procesamiento del Lenguaje Natural 71 (2023) 227–245.

[6] K. Sengupta, R. Maher, D. Groves, C. Olieman, Genbit: measure and mitigate gender bias in language datasets, Microsoft Journal of Applied Research 16 (2021) 63–71.

[7] T. Surana, T.-N. Ho, K. Tun, E. S. Chng, CASSI: Contextual and semantic structure-based interpolation augmentation for low-resource NER, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 9729–9742. URL: https://aclanthology.org/2023.findings-emnlp.651. doi:10.18653/v1/2023.findings-emnlp.651.

[8] B. T. Johns, M. N. Jones, Content matters: Measures of contextual diversity must consider semantic content, Journal of Memory and Language 123 (2022) 104313. URL: https://www.sciencedirect.com/

science/article/pii/S0749596X21000966. doi:https://doi.org/10.1016/j.jml.2021.104313.

[9] T. Heck, D. Meurers, On the relevance and learner dependence of co-text complexity for exercise difficulty, in: D. Alfter, E. Volodina, T. François, A. Jönsson, E. Rennes (Eds.), Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning, LiU Electronic Press, Tórshavn, Faroe Islands, 2023, pp. 71–84. URL: https://aclanthology.org/2023.nlp4call-1.9.

[10] T. Tyagi, C. G. Magdamo, A. Noori, Z. Li, X. Liu, M. Deodhar, Z. Hong, W. Ge, E. M. Ye, Y. han Sheu, H. Alabsi, L. Brenner, G. K. Robbins, S. Zafar, N. Benson, L. Moura, J. Hsu, A. Serrano-Pozo, D. Prokopenko, R. E. Tanzi, B. T. Hyman, D. Blacker, S. S. Mukerji, M. B. Westover, S. Das, Using deep learning to identify patients with cognitive impairment in electronic health records, in: Proceedings of Machine Learning Research ML4H, 2021. arXiv:2111.09115.

[11] J. Verschueren, Context and structure in a theory of pragmatics, Studies in Pragmatics 10 (2008) 14–24.

[12] T. Lai, H. Ji, T. Bui, Q. H. Tran, F. Dernoncourt, W. Chang, A context-dependent gated module for incorporating symbolic semantics into event coreference resolution, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3491–3499. URL: https://aclanthology.org/2021.naacl-main.274. doi:10.18653/v1/2021.naacl-main.274.

[13] L. Tamine, M. Daoud, Evaluation in contextual information retrieval: Foundations and recent advances within the challenges of context dynamicity and data privacy, ACM Comput. Surv. 51 (2018). URL: https://doi.org/10.1145/3204940. doi:10.1145/3204940.

[14] C. Hadiwinoto, H. T. Ng, W. C. Gan, Improved word sense disambiguation using pre-trained contextualized word representations, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5297–5306. URL: https://aclanthology.org/D19-1533. doi:10.18653/v1/D19-1533.

[15] D. Su, M. Patwary, S. Prabhumoye, P. Xu, R. Prenger, M. Shoeybi, P. Fung, A. Anandkumar, B. Catanzaro, Context generation improves open domain question answering, in: Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 793–808. URL: https://aclanthology.org/2023.findings-eacl.60. doi:10.18653/v1/2023.findings-eacl.60.

[16] C. Strathearn, D. Gkatzia, Task2Dial dataset: A novel dataset for commonsense-enhanced task-based dialogue grounded in documents, in: Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021), Association for Computational Linguistics, Trento, Italy, 2021, pp. 242–251. URL: https://aclanthology.org/2021.icnlsp-1.28.

[17] D. Ghosal, S. Shen, N. Majumder, R. Mihalcea, S. Poria, CICERO: A dataset for contextualized commonsense inference in dialogues, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5010–5028. URL: https://aclanthology.org/2022.acl-long.344. doi:10.18653/v1/2022.acl-long.344.

[18] R. Finkbeiner, J. Meibauer, P. B. Schumacher, What is a context? Linguistic approaches and challenges, volume 196 of *Linguistik aktuell = linguistics today*, John Benjamins Pub. Co., Amsterdam, 2012.

[19] G. Hollis, Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability, Journal of Memory and Language 114 (2020) 104146. URL: https://www.sciencedirect.com/science/article/pii/S0749596X20300607. doi:https://doi.org/10.1016/j.jml.2020.104146.

[20] G. Ferrari, Types of contexts and their role in multimodal communication, Computational Intelligence 13 (1997) 414–426.

[21] S. Castilho, J. L. Cavalheiro Camargo, M. Menezes, A. Way, DELA corpus - a document-level corpus annotated with context-related issues, in: Proceedings of the Sixth Conference on Machine Translation, Association for Computational Linguistics, Online, 2021, pp. 566–577. URL: https://aclanthology.org/2021.wmt-1.63.

[22] T. Udagawa, A. Aizawa, A natural language corpus of common grounding under continuous and partially-observable context, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, 2019. URL: https://doi.org/10.1609/aaai.v33i01.33017120. doi:10.1609/aaai.v33i01.33017120.

[23] B. Y. Lin, W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, X. Ren, CommonGen: A constrained text generation challenge for generative commonsense reasoning, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1823–1840. URL: https://aclanthology.org/2020.findings-emnlp.165. doi:10.18653/v1/

2020.findings-emnlp.165.

[24] F. Carlsson, J. Öhman, F. Liu, S. Verlinden, J. Nivre, M. Sahlgren, Fine-grained controllable text generation using non-residual prompting, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6837–6857. URL: https://aclanthology.org/2022.acl-long.471. doi:10.18653/v1/2022.acl-long.471.

[25] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, R. Xin, Free Dolly: Introducing the world's first truly open instruction-tuned LLM, 2023. URL: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[26] E. Sanchez-Bayona, R. Agerri, Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection, in: Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 228–240. URL: https://aclanthology.org/2022.conll-1.16. doi:10.18653/v1/2022.conll-1.16.

[27] V. Kovatchev, M. Taulé, InferES : A natural language inference corpus for Spanish featuring negation-based contrastive and adversarial examples, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3873–3884. URL: https://aclanthology.org/2022.coling-1.340.

[28] W. G. Reijnierse, C. Burgers, T. Krennmayr, G. Steen, The role of co-text in the analysis of potentially deliberate metaphor, in: Drawing Attention to Metaphor: Case studies across time periods, cultures and modalities, John Benjamins Publishing Company, 2020, pp. 15–38.

[29] J. Manyika, An overview of Bard: An early experiment with generative AI, Technical Report, Tech. rep., Technical report, Google AI, 2023. URL: https://ai.google/static/documents/google-about-bard.pdf.

[30] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., LLaMA: Open and efficient foundation language models, Computing Research Repository, arXiv:2302.13971. Version 1 (2023).

[31] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[32] E. G. Bard, D. Robertson, A. Sorace, Magnitude estimation of linguistic acceptability, Language 72 (1996) 32–68. URL: http://www.jstor.org/stable/416793.

[33] J. Novikova, O. Dušek, V. Rieser, RankME: Reliable human ratings for natural language generation, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 72–78. URL: https://aclanthology.org/N18-2012. doi:10.18653/v1/N18-2012.

[34] A. Turpin, F. Scholer, S. Mizzaro, E. Maddalena, The benefits of magnitude estimation relevance assessments for information retrieval evaluation, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 565–574. URL: https://doi.org/10.1145/2766462.2767760. doi:10.1145/2766462.2767760.

[35] S. Santhanam, S. Shaikh, Understanding the impact of experiment design for evaluating dialogue system output, in: Proceedings of the The Fourth Widening Natural Language Processing Workshop, Association for Computational Linguistics, Seattle, USA, 2020, pp. 124–127. URL: https://aclanthology.org/2020.winlp-1.33. doi:10.18653/v1/2020.winlp-1.33.

[36] R. Doust, P. Piwek, A model of suspense for narrative generation, in: Proceedings of the 10th International Conference on Natural Language Generation, Association for Computational Linguistics, Santiago de Compostela, Spain, 2017, pp. 178–187. URL: https://aclanthology.org/W17-3527. doi:10.18653/v1/W17-3527.

[37] S. Pita-Fernández, Determinación del tamaño muestral, Cuadernos de atención primaria 3 (1996) 138–141.

[38] C. Barros, M. Vicente, E. Lloret, To what extent does content selection affect surface realization in the context of headline generation?, Computer Speech & Language 67 (2021) 101179. URL: https://www.sciencedirect.com/science/article/pii/S0885230820301121. doi:https://doi.org/10.1016/j.csl.2020.101179.

[39] Y. G. Vázquez, A. F. Orquín, A. M. Guijarro, S. V. Pérez, Integración de recursos semánticos basados en WordNet, Procesamiento del lenguaje natural 45 (2010) 161–168.

[40] A. Siddharthan, N. Katsos, Offline sentence processing measures for testing readability with users, in: Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader

populations, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 17–24. URL: https://aclanthology.org/W12-2203.

[41] M. R. Costa-jussà, An analysis of gender bias studies in natural language processing, Nature Machine Intelligence 1 (2019) 495–496.

[42] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 632–642. URL: http://aclanthology.lst.uni-saarland.de/D15-1075.pdf.