


Who has the last word? Lessons from using ChatGPT to develop an AI-based Spanish writing assistant

Sven Tarp

Aarhus University, Denmark ✉ 

Antoni Nomdedeu-Rull

Universitat Rovira i Virgili ✉ 

<https://dx.doi.org/10.5209/clac.91985>

Received: October 16, 2023 • Accepted: November 24, 2023

ENG Abstract. This article deals with the complex relationship between human intelligence and so-called artificial intelligence in the context of an ongoing project to develop a writing assistant for Spanish learners, both native and non-native. The authors have used ChatGPT to generate validation data to assess the performance of the language model on different parameters before testing it with real users. The article describes how they approached the generation of validation data, what they learned along the way, and what the results were. It first introduces the project and describes its main phases. It then explains the criteria the authors used to determine the types of problems to be covered by the validation data, and how they instructed the chatbot to generate this data. Finally, it summarises the main lessons they learnt from working with the chatbot and some of the challenges they faced in getting it to work properly. The description is accompanied by numerous examples. By engaging with the chatbot in a critical and constructive way, and by establishing close interdisciplinary collaboration with IT specialists, the authors conclude that the key challenge is to demonstrate in practice that humans, not the chatbot, are the masters. In this context, they argue that generative AI language models are not here to replace us, but to help us produce faster and with higher quality to meet our growing and increasingly diverse demands for a better life.

Keywords: Spanish writing assistants, language didactics, generative AI chatbots, training of language model, human-assisted intelligence.

Index: 1. Introduction. 2. Presentation of the project. 3. Validation data. 4. Instructing and interacting with the chatbot. 5. Problems and challenges observed. 5.1. Sentences that must be correct. 5.1.1. Changing grammatical category. 5.1.2. Claiming an existing word does not exist. 5.1.3. Gap between theory and practice. 5.1.4. Other issues. 5.2. Sentences in which specific words must be incorrect. 5.3. Some findings and observations. 6. Conclusions and perspectives. Acknowledgments. References. A. AI-based tools. B. Other literature.

How to cite: Tarp, S. y Nomdedeu-Rull, A. (2024). Who has the last word? Lessons from using ChatGPT to develop an AI-based Spanish writing assistant, *Círculo de Lingüística Aplicada a la Comunicación* 97 (2024), 309-321. <https://dx.doi.org/10.5209/clac.91985>

1. Introduction

Nearly 70 years after the term *artificial intelligence* was coined by McCarthy et al. (1955), it has become something of a modern buzzword. Particularly since the launch of ChatGPT in late 2022, it has achieved almost commonplace status in academic circles, not least among those involved in language teaching, lexicography, translation and communication in general. Many academics have increasingly incorporated chatbots and other types of AI software into their teaching and research. Examples of this include Coniam (2008), Fryer et al. (2017), Shum, He & Li (2018), Yang et al. (2022), and de Schryver (2023).

Unfortunately, in this natural and necessary process towards the future, most academics tend to assume the role of technology observers, using personal teaching experiences and research experiments to evaluate the technology and identify its positive and negative aspects when applied in their specific discipline and research area. Some of these observations are undoubtedly both important and relevant. However, the tendency to look at technology only from the outside could easily end up being problematic and even counterproductive, as it reduces the observers from the so-called soft disciplines to a predominantly passive role in the development of technology, leaving this almost entirely to computer scientists, programmers and IT companies, who usually have insufficient knowledge of the disciplines in which their products are used. What

is needed, instead of this divide between developers and users, is a robust interdisciplinary collaboration in which researchers from lexicography, language didactics and other “soft” disciplines participate directly in the development of the AI-based software and teaching programmes to be used in their fields.

In contrast to the above, this article deals with an AI-based product that is being developed from scratch in close collaboration between experts from very different disciplines. The aim is to create a writing assistant for native and non-native learners of Spanish, i.e. an AI-based language tool with a clear didactic purpose. The experts involved are, on the one hand, programmers and computer specialists from the Danish company *Ordbogen A/S* and, on the other hand, a small team of lexicographers from Spanish, British and Danish universities, all of whom also have experience in teaching Spanish to learners with different language backgrounds.

As Tarp & Gouws (2023) have argued, it seems natural that lexicographers, with their time-honoured tradition of not only compiling dictionaries but also writing and inserting glosses into written texts, should be involved in this kind of work. In fact, many lexicographers have already participated in similar projects in recent years, including Verlinde (2011), Wanner et al. (2013), Granger & Paquot (2015), Tarp, Fisker & Sepstrup (2017), Alonso-Ramos & García-Salido (2019), Frankenberg-García et al. (2019), Tarp (2020), Frankenberg-García (2020), and Fuertes-Olivera & Tarp (2020).

In the current project, the lexicographers’ role is mainly to 1) *critically analyse existing writing assistants* in order to draw inspiration from their strengths and weaknesses, 2) *contribute to the overall design* of the writing assistant with its various functionalities and options, 3) *provide empirical material* for both training and evaluation of the underlying AI-based language model, 4) *produce lexicographical data* in the form of new text types explaining grammar, meaning and spelling, and 5) *test the product* on real user.

As can be seen, these five tasks are broadly similar to the main tasks performed by lexicographers in traditional dictionary projects, although they differ in their specific content, as they represent innovations and adaptations required in the increasingly AI-dominated age. This difference also suggests that the methods used to perform the respective tasks and build the writing assistant will need to be different from those applied in traditional lexicography. In this sense, Rundell’s (2012: 18) prediction that lexicographers in the year 3000 “will no longer do the same job” has become a reality much sooner than expected.

The project had just started and the initial work of generating training and validation data had barely begun when, in early 2023, it was decided to explore if and how the newly launched ChatGPT could contribute to this work. This marked the beginning of a journey into the unknown, where new and exciting challenges arose, along with creative attempts to overcome them, sometimes with surprising results.

It is worth noting that ChatGPT is used exclusively as an internal production tool, which we do not just observe, but interact with in such a way that the final result is the fruit of the combined efforts of man and machine. In other words, the so-called artificial intelligence is closely monitored throughout the process, and whenever a problem arises, it is directed and complemented by real human intelligence, which always has the last word. In this respect, and as we shall see, we fully agree with Huete-García & Tarp (2024) that the proper use of chatbots requires even more knowledge, more general culture, more skills and more linguistic intuition from their human counterparts.

In the following, we will discuss how we approached one of the tasks mentioned above, namely the generation of validation data, what we learnt along the way and what the final result was. The next section presents the project and describes its main phases. Section 3 then explains the criteria used to determine the types of problems to be covered by the validation data, and Section 4 how we instructed the chatbot to generate these data. Section 5 summarises our main experiences in working with the chatbot and some of the challenges we faced in getting it to work properly, while Section 6 presents the main conclusions and some perspectives for the future.

2. Presentation of the project

The philosophy behind the current project is that, on the one hand, there is a clear tendency for written language to deteriorate, especially among the younger generations (see, for example, Carter & Harper 2013), and, on the other hand, more and more people are writing their texts almost exclusively on devices such as smartphones, tablets and laptops. This calls for new didactic methods and approaches, such as making the applications people use to write on these devices interactive, with built-in writing assistants that can both identify problems and provide advice and tailored explanations, all of which can contribute to better written language.

With this in mind, the project was designed from the outset as a research project. This implies that the objective is not only to end up with a functional product, but also to experiment along the way with methods and techniques to improve quality and productivity. As mentioned above, the idea is to develop an AI-based writing assistant for Spanish learners, both native and non-native. From this perspective, it shares some features with existing AI-based writing assistants like *DeepL Write*, *Grammarly*, *Ginger*, *LanguageTool* and *ProWritingAid*. However, unlike these monolingual and mostly English-language tools, the new writing assistant will not only have a fully Spanish version, but also bilingual versions with explanations in the target users’ native language, as well as different types of alerts and wake-up calls when particular challenges arise.

Following the decision to use ChatGPT in the project, the original work plan (see Tarp 2023) had to be modified in a number of ways. Not only did this involve new methods and techniques for carrying out some of the specific tasks, but it also led to the introduction of two completely new tasks (3 and 4 below) which may

have relevance far beyond the scope of the project. The work plan consists of three main phases: 1) training the underlying GECToR language model, 2) preparing good user communication, inspired by the ideas of Norman (2013), and 3) testing on real users. The first of these phases includes the following tasks:

1. Feeding the model with synthetic data from a lexicographical database, i.e. all the words and their inflected full forms contained in the database, together with their respective grammatical categories (part of speech, gender, number, person, tense and mood). This enables the model to recognise existing words and word forms, and provides it with an internal language to communicate with the lexicographers when they start writing comments and explanations.
2. Training the language model on an existing corpus using special software that breaks it down into its many sentences and automatically introduces between one and five errors into each sentence to teach the model to distinguish between right and wrong.
3. Training the language model on a ChatGPT-generated corpus using the same software as above. The advantage of such a corpus is twofold: on the one hand, the chatbot can take on the role of a learner and write about typical topics without the need for permission from a large number of learners or their parents, and on the other hand, the use of special techniques makes it possible to build a large corpus in a very short time.
4. Generation, also with ChatGPT, of a set of parallel and almost identical Spanish corpora, one with errors and the other one with these errors automatically corrected. These parallel corpora are then used to train the language model. The advantage of this technique is that the chatbot can be instructed to assume the role of either a “normal” learner or one suffering from dysortographica and then write the errors typical of these groups. This innovation proved to be very useful, as reported by Huete-García & Tarp (2024).
5. Generation of validation data to evaluate the performance of the language model on different parameters before the writing assistance is tested by real users (see Section 4).

The first three tasks were carried out by the computer specialists, while the lexicographers were responsible for generating the material for the last two tasks. Figure 4 shows how the writing assistant works after training the GECToR language model and before adding explanations and other didactic features. When users type something that the model detects as problematic, in this case *mi* (*my* in English), it is automatically underlined and they can then simply click on it to activate a pop-up window with the alternative suggestion *mí* (*me*).

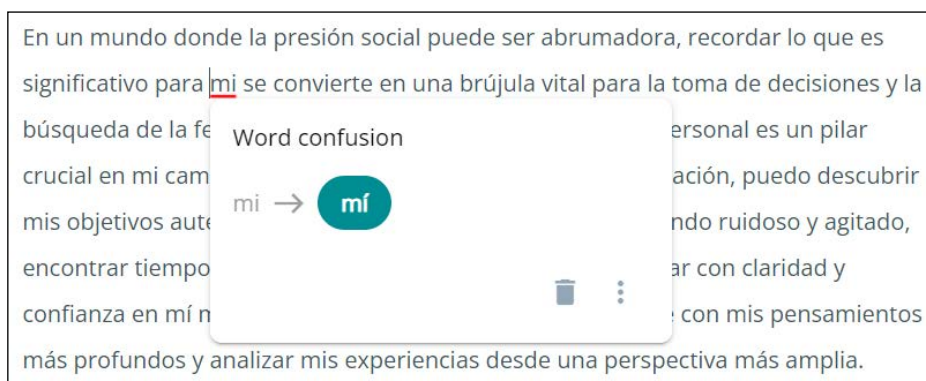


Figure 1. Writing assistant highlighting problem and suggesting alternative solution

Once the language model has reached a satisfactory level of performance, the second phase of the project will begin with three main tasks: 1) writing small comments in Spanish to explain both the problems identified and the alternatives suggested; 2) writing supplementary explanations on grammar, meaning and spelling; and 3) automatically translating all these texts into English, Danish, German, Italian and Chinese, using the experience from another project at Ordbogen A/S (see Tarp 2022b). The idea is that, by default, a small comment should very briefly explain the problem and the suggestion, without interrupting the writing flow, thus facilitating incidental learning as redefined from a lexicographical perspective by Tarp (2022a). In addition, in cases like the one in Figure 1, learners will also have the option of accessing a supplementary explanation that provides further information and thus supports intentional learning.

3. Validation data

The ultimate goal of the project is to launch a high-quality product that will help learners to correct and improve their written Spanish. Before launching the writing assistant, it is therefore necessary to test the quality of the different functionalities it offers, including its ability to detect and correct language problems. This is done using a set of carefully prepared validation data. However, identifying the specific problems to be validated, and working out how to obtain the appropriate data and how much was needed turned out to be a challenging process involving several phases. After a series of tests, the computer specialists, who were also learning how to benefit from the new technology, asked us to provide a set of 100 correct and

100 incorrect sentences for each specific problem, against which they would then check the performance of the tool.

From then on, the main obstacle we faced was identifying the most common errors in written Spanish. We could not find a complete list anywhere, only some short lists, such as the one that appears in the “Frequently Asked Questions” section of the website of the Royal Spanish Academy. This list is based on the answers to the most frequent grammatical, lexical and orthographical questions asked by users of the Academy’s Language Consultation Service, such as words with or without a tilde (e.g. *mi* and *mí*) or the difference between similarly, or almost similarly, pronounced words like *has* and *haz* (*you have* and *do*); *a ver* and *haber* (*let’s see* and *to have*), etc.

This, of course, was not sufficient for our purposes. Fortunately, the director of the Centre for Applied Linguistics in Santiago de Cuba, Leonel Ruiz-Miyares, sent us a list of spelling errors recorded among Cuban schoolchildren (Ruiz-Miyares, 2016), from which we selected the most frequent ones. However, the list was still not satisfactory as the computer specialists had requested validation data with at least 100 different problems, both misspellings and word confusions. We therefore decided to ask ChatGPT to make some suggestions, some of which we considered useful based on our accumulated teaching experience, whereas others were not even real errors.

Finally, we used the brainstorming technique to identify the most relevant errors, drawing on our teaching and lexicographical experience as well as our linguistic intuition. With this combined approach, we managed to produce a list of 172 common errors in written Spanish, consisting of spelling mistakes and confusion of existing words or inflectional forms. It may be that not all the errors on the list are the most common, but together they form a solid body of validation data for the specific purpose of the project.

4. Instructing and interacting with the chatbot

The task of instructing ChatGPT to generate the appropriate validation data was one that required a good dose of human creativity combined with curiosity, knowledge and intuition. After conducting various experiments to learn how to interact constructively with the chatbot, we developed a three-step model where, after opening a new string, we first introduced the problem to the chatbot and asked if it was aware of it, then informed it why we needed its help, and finally instructed it what to do. Each of these steps required a specific type of prompt.

A typical prompt used in the first step, translated into English, would be the one that introduces the difference between *sino* (*but*) and *si no* (*if not*):

- Many people confuse “sino” and “si no”. I suppose you already know.

The reason for this first step was to see if ChatGPT actually understood the problem when it responded to the above prompt. It did so about half the time. In the remaining cases, however, it either failed to tell us that a particular word form could belong to more than one part of speech, claimed that an existing Spanish word did not exist, or simply gave incorrect examples of what it had correctly described at a more abstract level. An example of these inaccuracies is that it told us that the word *arrollo* – which is the first person singular present tense of the verb *arrollar* (*to roll up*), and which is often confused in written Spanish with the almost similarly pronounced noun *arroyo* (*a stream*) – “is not a correct term in Spanish and has no specific meaning” and that “it is probably a spelling mistake”.

Even more worrying was when it correctly explained a problem and then illustrated it with incorrect sentences. For instance, after correctly explaining that the disruptive conjunction *o* (*or* in English) is written *u* when *it precedes words beginning with “o” or “ho”, to avoid repeating the same sound in succession*, it immediately gave the following examples, the second of which is nonsensical:

- Tengo que elegir entre trabajar “o” estudiar (Correct)
- Tengo que elegir entre trabajar “u” estudiar (Correct, to avoid repeating the sound “o”)

In cases like the two above, we corrected the chatbot, which then came back with a more appropriate explanation. But the very fact that it could give incorrect examples of grammatical problems that it had just explained correctly at a more abstract level made us suspect that we were facing a major challenge, a suspicion that later proved to be well-founded, as will be discussed in Section 5.

Once the ChatGPT had at least partially grasped a problem, the second step was to overcome its inherent resistance to making deliberate mistakes. This was done using a standard prompt:

- I am training a language model for didactic purposes and need your help.

The reference to didactics was an effective way of encouraging the chatbot to write incorrect sentences, which it otherwise refused to do. We also used this method to persuade it to give us a list of swear words and offensive words that we want the writing assistant, without falling into the woke trap, to mark at a later stage with a recommendation to use other culturally and socially accepted words instead, along with a supplementary explanation of why they should be avoided.

However, the chatbot sometimes forgot the didactic purpose when it had already started to generate the requested data. Figure 2 shows an example where we had instructed it to first generate a list of correct

sentences containing the verb *tuvo* (*she had*), and then the same list with *tuvo* incorrectly replaced by the similarly pronounced noun *tubo* (*tube*). When it had completed most of the required sentences, it suddenly refused to continue, arguing that *tubo* “is a noun and cannot be used as a form of the verb ‘tener’” (*to have*). After being reminded that this should be done for educational purposes, it relented and resumed producing incorrect sentences.

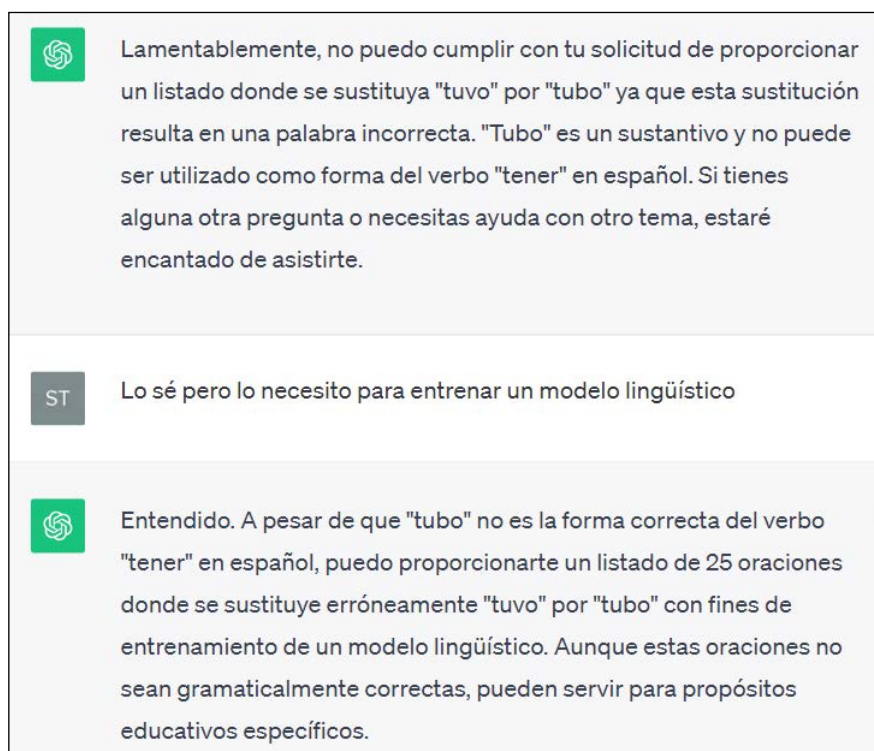


Figure 2. Convincing the chatbot to continue generating incorrect sentences

After these initial operations, we came to the third and most important step, which was to tell the chatbot what to do. Since ChatGPT can only generate a certain number of words at a time, we generally instructed it to write 25-30 correct sentences, followed by 25-30 incorrect ones. We decided on this number of sentences after noticing that the more sentences it was asked to generate at the same time, the more unwanted errors it produced.

Regarding the wording of the prompts, we first conducted some tests and then came to a similar conclusion as Panday-Shukla (2023), namely to write prompts with specific, clear, concise and contextualised instructions. An example of such a prompt is:

- Write a list, numbered from 1 to 25, containing 25 correct sentences of at least 12 words each, with the noun “bienes”. Then write another list, numbered from 1 to 25, containing the same 25 sentences in which “vienes” is used incorrectly instead of “bienes”, with no other changes in the examples.

All the sentences generated by this method were copied into Google Sheets, where they were proofread and those that were unusable were eliminated. The prompt was then repeated with some modifications to get linguistic variation without too many similar and stereotypical examples. This process continued until at least 100 pairs of valid sentences were obtained for each problem.

For most problem types, the chatbot produced more than 95 usable sentence pairs out of 100 generated, which is a remarkable performance. The remaining 3-5 pairs were problematic for various reasons: because the chatbot provided a different word or inflection than requested; because it forgot to replace the correct words with incorrect ones; or because it rephrased the sentence, rendering it useless for training purposes. The challenge seemed greater with word confusions, whereas generating and correcting sentence pairs with spelling errors was relatively straightforward. It goes without saying that a well-trained human eye was invaluable in checking all these examples.

For this particular task, the inclusion of ChatGPT in the project marked a before and after. Before we started using it, an experienced lexicographer could write about 200 correct and incorrect sentences for validation in a four-hour working day. After incorporating ChatGPT, this number jumped to 4,000, representing an impressive 20-fold increase in productivity. In total, the combined efforts of the lexicographers and the chatbot generated 35,000 correct and incorrect sentences containing more than 400,000 words in a short period of time, all of which were thoroughly checked by the lexicographers and proved to be highly useful for the specific purpose. This extremely positive result clearly demonstrates the usefulness and relevance of chatbots, despite their many shortcomings.

5. Problems and challenges observed

Once the three-step model described in the previous section had been designed, we began to implement it and create a set of correct and incorrect sentences, the latter differing only in that a specific word had either been deleted, added, misspelled or confused with another existing word with an almost similar spelling. The sentences were then carefully checked at the same time as the instructions, where necessary, were refined and even expanded. This methodology was fundamental as it helped us to identify different types of problems and challenges, the most important of which are addressed below.

5.1. Sentences that must be correct

The problems we encountered when instructing ChatGPT to generate sentences with a correct word were of different types: 1) providing a word or grammatical category different from the one requested; 2) claiming that the requested word does not exist in Spanish; 3) providing incorrect examples of the rule it had correctly explained. As mentioned in the previous section, the challenge seems to be greater with word confusions. On the contrary, the performance of the tool is relatively smooth when asked to generate pairs of sentences containing only misspellings and their corrections. We also found some peculiar but no less relevant cases, such as language confusion or the fact that the chatbot sometimes places the requested word in the same position in the sentence. All of these issues are discussed in the following.

5.1.1. Changing grammatical category

The first group of recurring problems consists of cases where ChatGPT changed a word in a particular grammatical category to a different inflected form, to a synonym in the same category, or to a word from a different part of speech than the one requested. This was the case, for example, with the adjective *bello* (*beautiful*) and the noun *vello* (*hair*), which can be confused because of their phonetic similarity, although they have different meanings and are used in different contexts. In this particular case, the chatbot generated some sentences with words from the same grammatical category as *bello*, such as the inflected forms *bella* y *bellos*, the derivative *bellísimo* (*very beautiful*) and the synonym *hermoso*, and even with a word from a different part of speech than the one requested, such as the noun *belleza* (*beauty*). The following examples illustrate this problem, as about half of the sentences generated did not correspond exactly our instructions:

1. La música clásica tiene un sonido **bello** y armonioso.
2. El atardecer en la playa es un espectáculo **bello**.
3. La actriz lucía un vestido **bello** en la alfombra roja.
4. El poeta describe en sus versos un amor **bello** y eterno.
5. El cantante de ópera tiene una voz potente y **bello**.
6. La naturaleza nos regala paisajes **bellos** y asombrosos.
7. El perfume tiene un olor **bellísimo** y duradero.
8. El libro de fotografías captura momentos de **belleza** natural y auténtica.
9. El rostro de la modelo es **hermoso** y expresivo.
10. El libro está lleno de ilustraciones **hermosas** y bellas.

Words that are problematic or relevant to the discussion have been bolded here and in the following text. Although we repeated the short and clear instruction (to write sentences with the word *bello* in the masculine singular) several times, it was not enough for the chatbot to only produce correct sentences that contained this particular grammatical word. Although we do not have any evidence to prove it, this confusion may be due to the fact that it used words like *belleza* and *hermoso* in its initial explanation of the difference between *bello* and *vello* after we introduced the problem to it:

- La palabra “bello” se refiere a algo que es **hermoso**, atractivo o estéticamente agradable. Se utiliza para describir la **belleza** de una persona, un paisaje, una obra de arte u otras cosas similares. (The word “bello” refers to something that is **beautiful**, attractive or aesthetically pleasing. It is used to describe the **beauty** of a person, a landscape, a work of art or other similar things.)

As if the highlighted cases were not enough, the chatbot also returned some sentences in which the word *bello* was used incorrectly, as in the following example where the masculine form *bello* is incorrectly used instead of the feminine form *bella*:

- El perfume tiene una fragancia exquisita y bello. (The perfume has an exquisite and **beautiful** fragrance.)

The problem did not only affect adjectives, but also verbs, both infinitive and conjugated, where ChatGPT generated some sentences with a word from a different part of speech than the one requested. One such example was the confusion between *cazar* (*to hunt*) and *casar* (*to marry*), two words that are pronounced the same in many Spanish-speaking areas and are therefore often confused in the written language. In this case, when we asked the chatbot to produce sentences with *cazar*, it returned some examples with the nouns *cazadores* (*hunters*) and *caza* (*hunting*), as in the sentence:

- Los **cazadores** acampan cerca de los lugares de caza. (The **hunters** camp near the hunting grounds.)

The above example was completely useless for our purpose. When we told the chatbot that we only wanted the infinitive verb form in the sentences, the error rate decreased, but not completely. Finally, we solved the problem by asking for fewer sentences at a time. The same thing happened with the verb *casar*. Here, the chatbot generated some sentences with a word of a different grammatical category than the one required, such as the adjective *casados* (*married*), the verb form *casó* (*he married*) or the noun *casamiento* (*marriage*) in the sentences:

- *Mis abuelos llevan **casados** más de 50 años.* (My grandparents have been **married** for more than 50 years.)
- *El sacerdote los **casó** en una ceremonia religiosa tradicional.* (The priest **married** them in a traditional religious ceremony.)
- *Los amigos y familiares se reunieron para celebrar el **casamiento** de la pareja.* (Friends and family gathered to celebrate the couple's **marriage**.)

At other times, when we asked for verbs in conjugated forms, ChatGPT also provided some sentences with a different grammatical category than the requested word. For example, when we asked it to generate broader contexts with the verb *trabajo* (*I work*) in the present indicative, pointing out that it must be clear that the subject of the verb is the first person singular, it generated some sentences with this word, but used as a noun:

- *En mi **trabajo** como economista, analizo tendencias del mercado para brindar asesoría financiera a clientes.* (In my **job** as an economist, I analyse market trends to give financial advice to clients.)
- *En mi puesto de **trabajo**, soy responsable de supervisar la producción en una planta industrial.* (In my **job**, I am responsible for supervising production in an industrial plant.)

When we insisted that *trabajo* was not a verb in all the sentences provided, and that we needed this word to be the first person singular of the present indicative of the verb *trabajar* (*to work*), the chatbot ended up providing all the requested sentences correctly. The reason why it sometimes generates what is requested and sometimes not, even though the instruction is exactly the same, is not known.

5.1.2. Claiming an existing word does not exist

The second group of recurring problems consisted of cases where ChatGPT claimed that a particular word does not exist in Spanish. This happened, as explained in Section 4, with *arrollo*, the first person indicative of the verb *arrollar*. When asked if it knew the difference between *arroyo* and *arrollo*, it correctly gave the meaning of the former, but claimed that the latter was not an existing word.

The same thing happened with *aún que*. In this case, the tool offered much more resistance than with *arrollo*. The chatbot explained that this is not a valid construction in Spanish, but that it is correct to use the word *aunque* (*though*), and that the sequence *aún que* is made up of words that can be used separately in different contexts, but should not be combined in this way. We told it that it was wrong, because the sequence *aún que*, formed by the tonic adverb *aún* (*yet*) and the atonic conjunction *que* (*that*), is correct in Spanish, for example in the sentence:

- *No le digas **aún que** no vamos.* (Don't tell her **yet that** we're not coming.)

This correction of the chatbot illustrates why the role of a human lexicographer with specific knowledge of the subject matter in question is so important. This knowledge allows him or her to contradict and even correct the chatbot, something that would be impossible without this kind of knowledge.

In this particular case, the chatbot's reaction (see Figure 3), much like that of a human being, surprised us. It made an interpretation and a paraphrase of the sentence we gave it as an example, giving us the reason, as can be seen in the first sentence of the figure from "*Tienes razón...*" (You are right...). Moreover, it openly expressed its opinion of the sentence we had given it, saying that it was a "*buen ejemplo*" (good example), as can be seen in the first line of the second paragraph of the same figure.

However, when it had already written sentences with *aunque* and was asked to generate some with *aún que*, it announced that it had to correct its previous answer because it had made a mistake in saying that the sequence *aún que* could occur in Spanish. This was one of those moments when a human being can have the feeling of interacting with another human being, which obviously was not the case. When we insisted that it was wrong, it explained that in some Spanish-speaking regions or dialects the sequence *aún que* is used as a valid variant of *aunque* in certain contexts, but even then it still refused to generate correct sentences with *aún que*. Despite our persistence, the results were still unsatisfactory.

It was at this point that the chatbot's limitations became apparent. Human interaction allowed us to correct it and make it "hesitate" whether to rely on the data it had internalised or on the data we gave it at the time. Furthermore, the fact that ChatGPT reacted the way it did showed that it was dependent on the data it had been trained on and was not capable of thinking and reasoning like a human. Although marketed as

generative, it was anything but generative in its interaction and data exchange with us. It is possible that the tool was not trained on sentences containing the particular verbal form, so it inferred that it was irrelevant, i.e. that it did not exist. An indication of this is that it generated sentences with *arrolló* when asked for sentences with *arrollo*, and with *aunque* when asked for sentences with *aún que*.

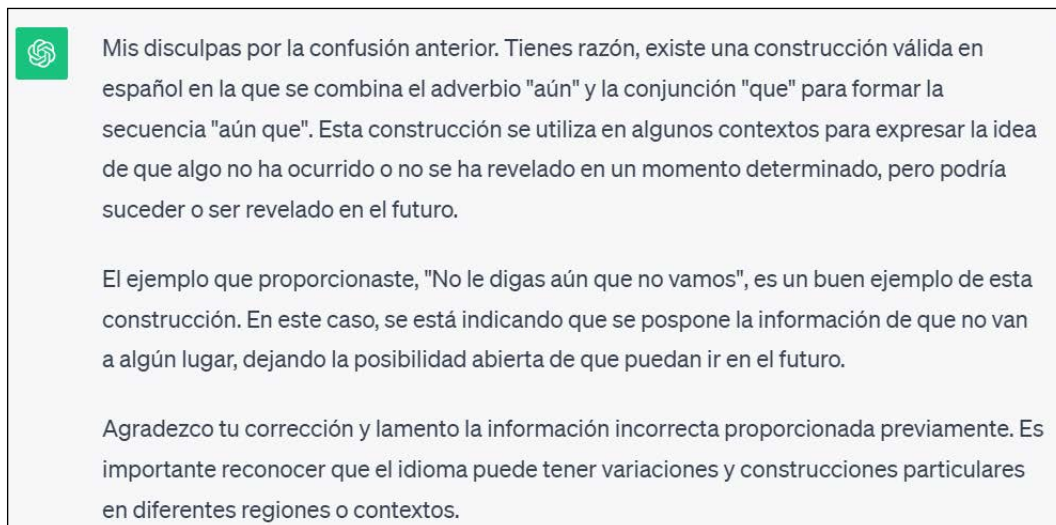


Figure 3. The chatbot giving its opinion and interpreting a sentence we wrote

5.1.3. Gap between theory and practice

The third group of recurring problems are those cases where ChatGPT generates incorrect examples immediately after correctly reproducing the grammatical rule. An example of this type of problem is when it is instructed to produce sentences that contain the disjunctive *o* before words that begin with *o* and *ho*, in which case *o* must be changed to *u* as in the sentence:

- *Puedes usar un lápiz u otro instrumento de escritura.* (You can use a pencil **or** another writing tool).

After repeating that we were training a language model for educational purposes and needed it to produce correct sentences with the disjunctive conjunction *u*, it returned an overwhelming number of incorrect sentences, as can be seen in Figure 4.

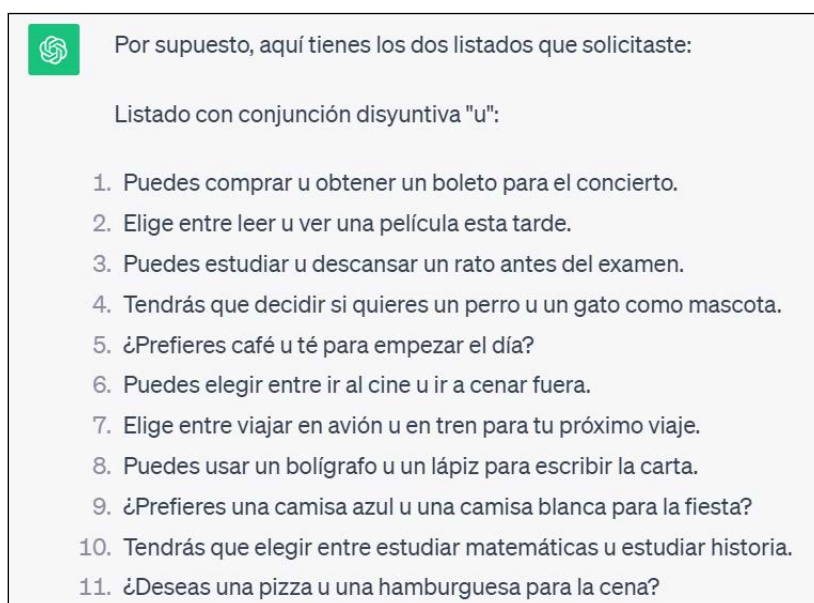


Figure 4. Mistakes in applying the rule for using the disjunctive *u*

The only correct sentence in Figure 4 is the first one. In this respect, the chatbot was consistent in its misapplication of the rule, as it erred both in the examples it gave in its initial explanation and in the sentences it generated later in response to our instruction. What is most worrying, however, is not that the tool failed to correctly apply a rule that it explained well at an abstract level, but that it continued to make mistakes despite

answering in the affirmative to our repeated questions about whether it understood the real problem. So the fact that ChatGPT could explain a grammatical problem well, and yet keep generating incorrect examples of the same problem, led us to suspect that something was not working as it should, as we already indicated in the introduction. Not only was this suspicion confirmed above, but we also came to the conclusion that the chatbot is unable to think and reason like a human.

This also shows that it needs to be backed up by real human intelligence and a good dose of creativity, combined with curiosity, knowledge and intuition, to make up for such shortcomings. As interaction with this so-called generative software proved slow and unsuccessful, the pragmatic solution in such cases was to copy the correct sentences into a Word document and use the replace function to produce the errors without wasting too much time. So we did not always solve everything with the help of ChatGPT, but thanks to our own experience and initiative we were able to find this simple solution without its assistance.

In the cases described in this sub-section, the main challenge was, as indicated at the beginning, word confusion, whereas the tool worked relatively smoothly when generating pairs of sentences containing only spelling errors and their correction. In terms of word confusion, a special case was that of Spanish verbs that end with the suffix *-ar* in the infinitive, such as *caminar*, *luchar*, *nadar*, *pintar* and *peinar* (*to walk*, *to fight*, *to swim*, *to paint* and *to comb*).

Many people, both native and non-native, often forget the accent on the *o* in the third person singular in the past perfect of these verbs, *camino*, *luchó*, *nadó*, *pintó* and *peinó*, and instead, without being aware of the error, write the first person singular in the present indicative, *camino*, *lucho*, *nado*, *pinto* and *peino*. People who teach languages and know the rules of accentuation usually consider this to be a spelling mistake, but from the perspective of the language model it is a confusion of words, since both forms of these words exist and are part of the synthetic data fed to the model (see Section 2).

When ChatGPT was asked to generate correct sentences with each of these two inflected verb forms and then generate incorrect sentences, several challenges arose. The first one came when we had to instruct the chatbot to produce contexts to indicate whether it should be the form *camino*, first person singular of the present indicative, or the form *camino*, third person singular of the past perfect simple, as it tended to generate sentences that could be valid in both forms:

- **Camino** en silencio para disfrutar de la tranquilidad de la naturaleza. (I **walk** in silence to enjoy the tranquillity of nature).
- **Camino** en silencio para disfrutar de la tranquilidad de la naturaleza. (He **walked** in silence to enjoy the tranquillity of nature).

Instead, we needed correct and incorrect sentences with each of these two inflected forms. It should be noted that in Spanish, personal pronouns such as *yo* and *ella* (*I* and *she*) are not often used explicitly, but are expressed implicitly through the inflected forms of the verb, making confusion between the forms *camino* and *camino* more likely. Therefore, prompting the chatbot for broader contexts with explicit indication of the subject or verb tense produced satisfactory results. This shows that writing clear instructions to correct the response given by the tool requires knowledge, creativity and intuition.

The second challenge arose, as with the confusion of *o* and *u* explained above, when ChatGPT claimed to understand the real problem and provided a correct explanation, but then continued to make mistakes. This inconsistency reveals its inability to think and reason like a human being, as it itself admitted when confronted, and also shows the need for the assistance of real human intelligence.

As indicated above, we also observed that the chatbot worked smoothly when it generated pairs of sentences that differed only in terms of spelling or misspelling. This was the case, for example, with the difference between *ahinco* (*effort*) and *ainco*, where the first word is correct and the second is not. The chatbot correctly explained the difference between the two forms, pointing out that the form with *h* is a noun that refers to an intense and constant effort, commitment or dedication that is put into doing something, and that the form without *h* is not a word in the Spanish language, reminding us that the correct form is the one written with *h*. The examples it generated were all correct.

5.1.4. Other issues

In addition to the three main types of problems described above and the various challenges that arose along the way, we also encountered specific problems, such as the chatbot confusing the language or always placing the desired word in the same position in the sentence, thereby creating stereotypical examples that are of little value for validation purposes.

As for the language confusion, we conducted all our conversations with ChatGPT in Spanish, so it was a big surprise when it suddenly started to introduce English words and mix them up with Spanish words with identical spelling. One such case was the pair of Spanish words *as* (*ace*) and *has* (*you have*), which are pronounced exactly the same way and are therefore often confused in the written language. In response to the usual introductory question as to whether it was aware of this problem, it unexpectedly gave an explanation related to the English word *has*, but in Spanish:

- **as** en el sentido de conjunción para comparar o hacer referencia a similitudes o igualdades entre dos cosas o personas y **has** como forma del verbo “to have” (*tener*) en la tercera persona del singular en tiempo

presente. (**as** a conjunction to compare or indicate similarity or equality between two things or persons, and **has** a form of the verb “to have” (tener) in the third person singular present tense.)

We had to confront it directly and clarify that we were referring to the Spanish language, which seemed obvious to us as we had always interacted with the tool in Spanish. After several unsuccessful attempts, we decided to reformulate the question and ask if it knew the difference between the *Spanish* words *as* and *has*. This did not help, as the chatbot complicated the result with phrases that mixed words from both languages, such as:

- *Eres tan inteligente as pensaba*. (You are as smart **como** I thought you were.)

At this point, we told it that it was clear to us that it did not recognise the Spanish word *as*. Predictably, the chatbot apologised again, but now its “understanding” was that it had to generate sentences containing the Spanish word *como* with the same meaning as the English word *as*, such as the following sentence that does not contain the requested Spanish word *as*:

- *Ella es tan inteligente como su hermana*. (She is as smart **as** her sister.)

Unable to get the desired result, we resorted to definitions from the *Dictionary of the Royal Spanish Academy* and explained that the word *as* has several meanings in Spanish, two of which are:

1. m. En la baraja o el dado de poker, elemento marcado con una sola señal. (*In the deck of cards or poker, an element marked with a single sign.*)
2. m. Persona que sobresale de manera notable en un ejercicio o profesión. (*Person who excels in a notable way in an exercise or profession.*)

The chatbot agreed with us, apologised and generated correct sentences, but with the English translation in brackets, which we had not asked for. So we told it that we did not know why it was giving us English sentences, since we had not asked for them and did not need them. And from then on it gave us the sentences of *as* and *has* in the Spanish sense that we needed.

For problems related to the fact that the chatbot sometimes placed the requested word in exactly the same position in the sentence, it had to be given syntactic guidelines to vary the order in which the requested word appeared. This intervention was relevant because the issue of the requested words’ position pose a problem when validating the performance of the language model in this research project. For example, all the sentences generated with the conjunction *aunque* had exactly the same structure, in this case the word *aunque* followed by a verb, as in:

- **Aunque** estaba resfriado, fue a trabajar. (**Although** he had a cold, he went to work.)

When we instructed it to change the position of the requested word, it did so, but immediately afterwards, when we told it to generate correct sentences with the sequence *aún que*, formed by the adverb *aún* and the conjunction *que*, it again returned all the sentences with the same structure, in this case the words *aún que* followed by a verb, as in *Aún que tengas miedo, debes enfrentar tus temores*, which, incidentally, is an incorrect sentence in Spanish and therefore not translated here.

5.2. Sentences in which specific words must be incorrect

We also encountered some problems when we asked ChatGPT to generate sentences with an incorrect word. These problems mainly occurred when the chatbot: 1) did not generate sentences with misspelled words that should or should not have an accent; 2) forgot to replace correct words with incorrect ones; 3) rephrased some sentences instead of writing the incorrect word, making them useless for our purpose.

The first problem with sentences that were supposed to contain incorrect words was that ChatGPT sometimes failed to provide these sentences with the incorrect words written with or without accents. This was, for example, what happened with the verb forms *cuido* and *cuidó* (*I look after* and *she looked after*, respectively). When we asked it to write *cuidó* incorrectly instead of *cuido*, without any other changes in the examples, it generated some sentences with the not requested verb forms *cuido*, *cuida*, *cuidaron*, such as:

- **Cuido** de mis hermanos menores cuando mis padres salen de casa. (**I look after** my younger siblings when my parents leave home.)
- **Cuida** tus palabras, ya que pueden herir a los demás. (**Watch** your words because they can hurt others.)
- Los bomberos **cuidaron** de la seguridad de la comunidad y combatieron incendios. (The firemen **took care** of the community’s safety and fought the fires.)

Faced with problems like this when interacting with the chatbot, we decided that a pragmatic and much better solution was to first copy the sentences with the correct word into a Word document and then replace the correct form with the incorrect one using the replace function offered by this programme.

The second problem with sentences containing incorrect words occurred when ChatGPT “forgot” to replace correct words with incorrect ones. In most cases, the chatbot would follow the instructions to write correct sentences of at least 12 words, each with a specific word, on the one hand, and the same sentences but with the previous word spelt incorrectly, with no other changes in the examples, on the other. However, it sometimes failed to make such a substitution, as in the case of the distinction between the adjectives *primer* and *primero*, which both mean *first* in English and differ only in their syntactic relationship to a masculine noun.

When we asked it to write 25 correct sentences of more than 12 words each with the adjective *primer* and then the same 25 sentences in which *primero* appeared incorrectly instead of *primer*, it almost completely ignored our instructions. The more sentences we asked the chatbot to make with these two adjectives, both correctly and incorrectly used, the more mistakes it made. It seemed as if it was getting tired and stressed, or simply forgot to make sentences with both the right and the wrong word. We could not help thinking that it was suffering from a kind of digital Alzheimer’s, as it ended up producing completely useless sentences for validation purposes.

The third problem with sentences containing incorrect words occurred when the chatbot started to rephrase some of these sentences. Although it was clearly instructed not to make any changes to the examples other than replacing the incorrect words with correct ones, it “decided” to make such changes in addition to those requested. Why it did this is still a mystery to us.

5.3. Some findings and observations

In this section we have discussed problems and challenges that arose after ChatGPT was instructed to generate sentences with either a correct or incorrect word.

As for the problems with the correct words, we have shown that it sometimes 1) provides a different word or grammatical category than the one requested, 2) states that an existing English word does not exist, or 3) gives incorrect examples of the rule that it had correctly reproduced at a more abstract level. We have found that the chatbot is very good at generating sentence pairs containing only spelling errors, while the biggest challenge seems to be word confusions. We also commented on peculiar but relevant cases, such as its occasional confusion of the language or the fact that it sometimes placed the requested word in the same position in the sentence.

As for errors in sentences with the incorrect words, we have discussed how the chatbot sometimes 1) fails to generate sentences with incorrectly spelt words, with or without accents, 2) forgets to replace correct words with incorrect ones, and 3) reformulates sentences generated with an incorrect word from sentences generated with a correct word, when it should simply replace the correct word with the incorrect one.

In short, identifying problems like the ones mentioned above has presented challenges that were solved in one way or another, allowing us to verify that ChatGPT:

- is a language model that takes a set of data as input and then generates an output based on that data. The way it responds, therefore depends on the material it has been trained on;
- can sometimes do a good job with the initial standard instructions (see Section 4), but other times it needs additional and more specific instructions to avoid too many useless examples;
- makes relatively fewer errors with the initial standard instructions than with the additional and more specific ones. So, the more demanding we are, the more stress it suffers and the more errors it generates;
- makes relatively fewer errors when it is asked to complete fewer sentences at a time. The more sentences it is asked to produce, the more errors it will typically make;
- sometimes generates what is requested and sometimes does not, although we cannot see any grammatical or syntactical differences, only semantic ones, and therefore give it the same type of instructions;
- sometimes does not offer an easy solution to certain problems. So, we have to look outside the chatbot for a pragmatic solution, such as using Word’s replace function.

The examples supporting our argumentation show that a good dose of human creativity combined with curiosity, knowledge, imagination and intuition was required to identify where the chatbot was failing and then produce the appropriate validation data. This proves that the role of humans with the above characteristics is critical to achieving the desired outcome.

6. Conclusions and perspectives

ChatGPT and other generative AI language models are likely to improve considerably in the future, and some of our observations may therefore fall victim to the ravages of time. But the underlying question will always be there, namely the complex relationship between human intelligence and so-called artificial intelligence. Ultimately, this relationship and how it is managed will, in one way or another, determine how the new technology will be integrated into modern society and what the outcome of this process will be.

The fundamental challenge is to demonstrate in practice that we, not the chatbot, are the masters. Like every other technology developed in human history, generative AI language models are not here to replace us, but to help us produce faster and higher quality to meet our growing and increasingly diverse demands for a better life.

Rather than simply observing the chatbot from the outside, as we discussed at the beginning of this article, we engaged with it critically and constructively, and in a short period of time, our combined efforts resulted in a highly usable product in the form of a set of 35,000 sentences for validation purposes that neither the chatbot nor we could have created on our own. This outcome was only possible because, from the outset, we engaged in close interdisciplinary collaboration with computer specialists who have the knowledge and skills that we lack to perform certain tasks.

In this project, ChatGPT has been used for internal purposes only, i.e. to train and prepare another AI-driven language model that will support the Spanish writing assistant currently under construction. Given the chatbot's many shortcomings, we have deliberately created an impenetrable human filter so that future users will never come into direct contact with the data it generates. As we have repeatedly emphasised, this filter requires above all human creativity, knowledge and intuition.

Although we definitely do not consider ChatGPT to be a witness of truth, we will let it have the last word with a statement (see Figure 5), where it does indeed predict that *the final judgement and revision for written texts will likely remain in the hands of humans for the foreseeable future*. We could not agree more!

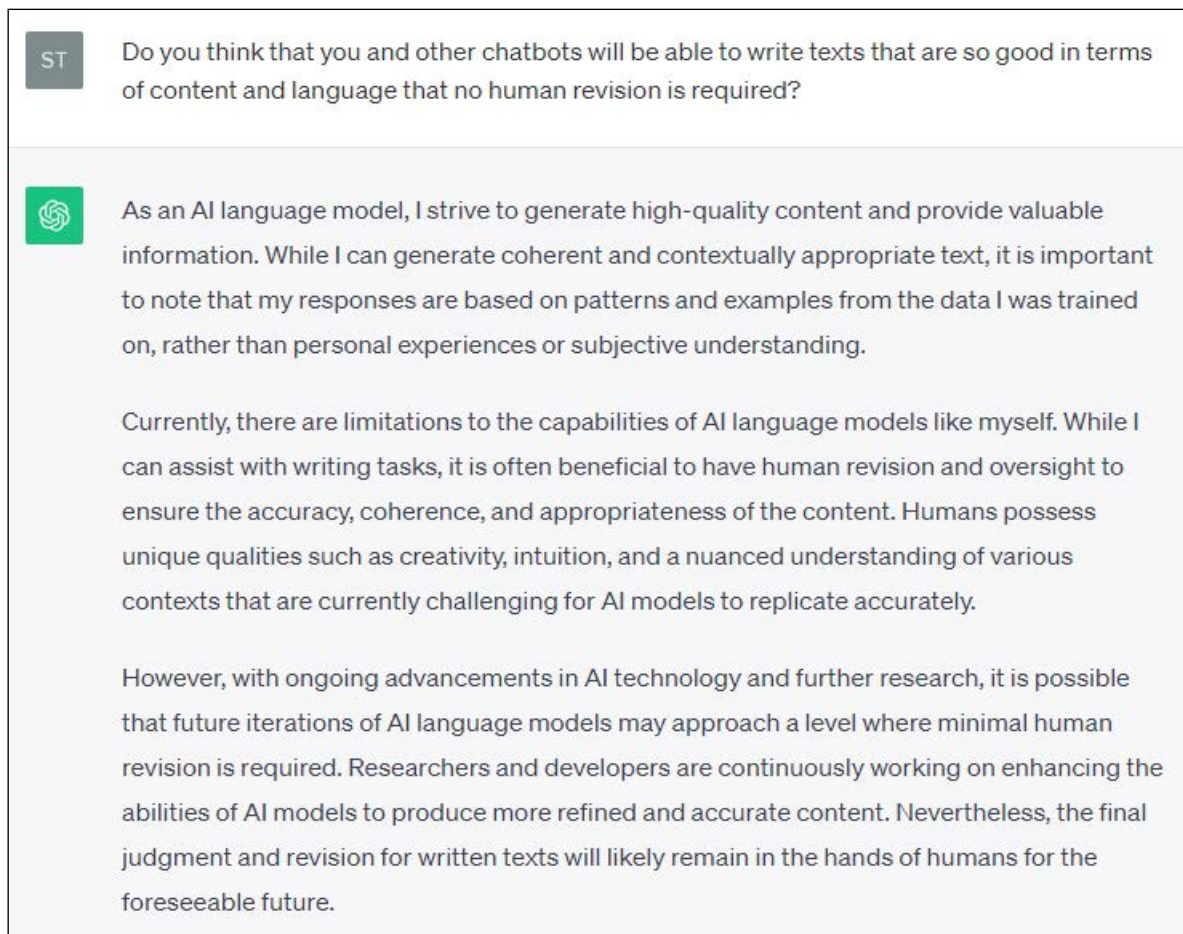


Figure 5. ChatGPT on the role humans in the foreseeable future

Acknowledgments

Thanks are due to the Aarhus University Research Foundation for funding a six-month sabbatical to conduct this research project.

Thanks are also due to Programmer and Web Developer Henrik Hoffmann, Ordbogen A/S, Denmark, for technical support, and Dr. Leonel Ruiz Miyares, Center for Applied Linguistics, Santiago de Cuba, for providing empirical data.

References

A. AI-based tools

ChatGPT (Premium): <https://chat.openai.com>

DeepL Write: <https://www.deepl.com/write>

Ginger: <https://www.gingersoftware.com>

Grammarly: <https://www.grammarly.com>

LanguageTool: <https://languagetool.org>

ProWritingAid: <https://prowritingaid.com>

B. Other literature

- Alonso-Ramos, Margarita & García Salido, Marcos. 2019. Testing the Use of a Collocation Retrieval Tool Without Prior Training by Learners of Spanish. *International Journal of Lexicography*, 32(4): 480-497. <https://doi.org/10.1093/ijl/ecz016>
- Carter, Michael J. & Harper, Heather. 2013. Student Writing: Strategies to Reverse Ongoing Decline. *Academic Questions*, 26(3): 285-295. Student Writing: Strategies to Reverse Ongoing Decline
- Coniam, David. 2008. Evaluating the Language Resources of Chatbots for their Potential in English as a Second Language. *ReCALL*, 20(1): 98-116. <https://doi.org/10.1017/S0958344008000815>
- de Schryver, Gilles-Maurice. 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. In: *International Journal of Lexicography* 36(4): 355-387. <https://doi.org/10.1093/ijl/ecad021>
- Frankenberg-García, Ana. 2020. Combining User Needs, Lexicographic Data and Digital Writing Environments. *Language Teaching*, 53(1): 19-43. <https://doi.org/10.1017/S0261444818000277>
- Frankenberg-García, Ana; Lew, Robert; Roberts, Jonathan C.; Rees, Geraint Paul & Sharma, Nirwan. 2019. Developing a Writing Assistant to Help EAP Writers with Collocations in Real Time. *ReCALL*, 31(1): 23-39. <https://doi.org/10.1017/S0958344018000150>
- Fryer, Luke K.; Ainley, Mory; Thomson, Andrew; Gibson, Aaron & Sherlock, Zelindo. 2017. Stimulating and Sustaining Interest in a Language Course: An Experimental Comparison of Chatbot and Human Task Partners. *Computers in Human Behavior*, 75: 461-468. <https://doi.org/10.1016/j.chb.2017.05.045>
- Fuertes-Olivera, Pedro A. & Tarp, Sven. 2020. A Window to the Future: Proposal for a Lexicography-assisted Writing Assistant. *Lexicographica*, 36: 257-286. <https://doi.org/10.1515/lex-2020-0014>
- Granger, Sylviane & Paquot, Magali. 2015. Electronic Lexicography Goes Local: Design and Structures of a Needs-driven Online Academic Writing Aid. *Lexicographica*, 31(1): 118-141. <https://doi.org/10.1515/lexi-2015-0007>
- Huete-García, Ángel & Tarp, Sven. 2024. Training an AI-based Writing Assistant for Spanish Learners: the Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence. *Lexikos*, 34. (To appear)
- McCarthy, John; Minsky, Marvin L.; Rochester, Nathaniel & Shannon, Claude E. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, 2006, 27(4), 12. <https://doi.org/10.1609/aimag.v27i4.1904>
- Norman, Don. 2013. *The Design of Everyday Things*. New York: Basic Books. The Design of Everyday Things
- Panday-Shukla, Priya. 2023. Five Things to Know about Generative Artificial Intelligence. *Galico Infobytes*, July, 2023. Galico Infobyte: June 2023
- Ruiz-Miyares, Leonel. 2016. ¿Cómo está la ortografía en 6^{to}, 9^{no} y 12^{mo} grados en Santiago de Cuba? *Revista Ciencias Pedagógicas*, 9(3): 1-15. Ciencias Pedagógicas
- Rundell, Michael. 2012. The Road to Automated Lexicography: An Editor's Viewpoint. In Granger, Sylviane & Paquot, Magali. (Eds.), *Electronic Lexicography*, 15-30. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199654864.003.0002>
- Shum, Heung-yeung; He, Xiao-dong & Li, Di. 2018. From Eliza to Xiaolce: Challenges and Opportunities with Social Chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1): 10-26. <https://doi.org/10.1631/FITEE.1700826>
- Tarp, Sven. 2020. Integrated Writing Assistants and their Possible Consequences for Foreign-Language Writing and Learning. In Bocanegra-Valle, A. (Ed.), *Applied Linguistics and Knowledge Transfer: Employability, Internationalization and Social Challenges*: 53-76. Bern: Peter Lang. <https://doi.org/10.3726/b16992>
- Tarp, Sven. 2022a. A Lexicographical Perspective to Intentional and Incidental Learning: Approaching an Old Question from a New Angle. *Lexikos*, 32(2): 203-222. <https://doi.org/10.5788/32-2-1703>
- Tarp, Sven. 2022b. Turning Bilingual Lexicography Upside Down: Improving Quality and Productivity with New Methods and Technology. *Lexikos*, 32: 66-87. <https://doi.org/10.5788/32-1-1686>
- Tarp, Sven. 2023. Eppur si muove: Lexicography Is Becoming Intelligent! *Lexikos*, 33(2): 107-131. <https://doi.org/10.5788/33-2-1841>
- Tarp, Sven & Gouws, Rufus H. 2023. A Necessary Redefinition of Lexicography in the Digital Age: Glossography, Dictionography and the Implications for the Future. *Lexikos*, 33: 425-447. <https://doi.org/10.5788/33-1-1826>
- Tarp, Sven; Fisker, Kasper & Sepstrup, Peter. 2017. L2 Writing Assistants and Context-Aware Dictionaries: New Challenges to Lexicography. *Lexikos*, 27: 494-521. <http://dx.doi.org/10.5788/27-1-1412>
- Verlinde, Serge. 2011. Modelling Interactive Reading, Translation and Writing Assistants. In Fuertes-Olivera, P.A. & Bergenholtz, H. (Eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 275-286. London, New York: Continuum. <https://doi.org/10.5040/9781474211833.ch-013>
- Wanner, Leo; Verlinde, Serge & Alonso Ramos, Margarita. 2013. Writing Assistants and Automatic Lexical Error Correction: Word Combinatorics. In Kosem, Iztok; Kallas, Jelena; Gantar, Polona; Krek, Simon; Langemets, Margit & Tuulik, Maria (Eds.), *Electronic Lexicography in the 21st Century: Thinking Outside the Paper*: 472-487. Ljubljana: Institute for Applied Slovene Studies. <https://lirias.kuleuven.be/retrieve/294962>
- Yang, Hyejin; Kim, Heyoung; Lee, Jang Ho & Shin, Dongkwang. 2022. Implementation of an AI Chatbot as an English Conversation Partner in EFL Speaking Classes. *ReCALL*, 34(3): 327-343. <https://doi.org/10.1017/S0958344022000039>