http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)

# Training an AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence

Ángel Huete-García, *University of Oxford, Faculty of Medieval and Modern Languages, United Kingdom (angel.huete-garcia@mod-langs.ox.ac.uk) (https://orcid.org/0000-0002-8098-7401)*
and
Sven Tarp, *Aarhus University, Denmark, and Stellenbosch University, South Africa (st@cc.au.dk) (https://orcid.org/0000-0003-1941-9082)*

**Abstract:** This article deals with the relationship between human and artificial intelligence in the context of an ongoing Spanish Writing Assistant project, where ChatGPT is used to assist in four key tasks related to either training the underlying language model or preparing future user communication. The project is an interdisciplinary collaboration between lexicographers with experience in language teaching and IT experts from a high-tech company. The article first describes the methodology of the overall project and the specific role of the lexicographers. It then discusses the three tasks in which the latter are directly involved: the construction of a set of two parallel Spanish corpora, one correct and the other with induced errors, the generation of validation material, and the writing of extended grammatical explanations for Spanish learners. Based on a large amount of empirical data, including 35,000 carefully reviewed sentences, the article details the different steps of the interaction between human and chatbot, as well as the experiences and reflections drawn from this process. It concludes that the two parts engage in very different types of relationships depending on the concrete task, and that human knowledge, culture, skills and language intuition are crucial for the chatbot to work properly.

**Keywords:** SPANISH WRITING ASSISTANTS, LANGUAGE LEARNING, CHATBOTS, HUMAN-ASSISTED INTELLIGENCE, TRAINING OF LANGUAGE MODEL, CORPUS BUILDING

**Opsomming: Die afrigting van 'n KI-gebaseerde skryfhulpmiddel vir Spaanse leerders: Die nut van kletsbotte en die onmisbaarheid van mensgesteunde intelligensie.** Hierdie artikel handel oor die verhouding tussen menslike en kunsmatige intelligensie binne die konteks van 'n lopende Spaanse skryfhulpmiddelprojek, waar ChatGPT vir vier sleuteltake verwant aan óf die afrigting van die onderliggende taalmodel óf die voorbereiding van toekomstige gebruikerskommunikasie aangewend word. Dit is 'n interdissiplinêre samewerkingsprojek tussen leksikograwe met ervaring in taalonderrig en IT-spesialiste van'n hoë-tegnologiemaatskappy. In die artikel word die metodologie van die algehele projek en die spesi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



fieke rol van die leksikograwe eers beskryf. Daarna word die drie take waarby die laasgenoemdes direk betrokke is, beskryf: die samestelling van 'n stel van twee parallelle Spaanse korpora, een korrek en die ander met ingevoerde foute, die generering van geldigheidsmateriaal, en die skryf van uitgebreide grammatikale verklarings vir Spaanse leerders. Gebaseer op 'n groot hoeveelheid empiriese data, insluitend 35,000 versigtig beoordeelde sinne, word die verskillende stappe van die interaksie tussen mens en kletsbot, sowel as die ervarings en refleksies wat uit die proses verkry is, uiteengesit. Daar word tot die gevolgtrekking gekom dat die twee dele, afhangende van die konkrete taak, in baie verskillende tipes verhoudings betrokke is, en dat menslike kennis, kultuur, vaardighede en taalintuïsie noodsaaklik is vir die kletsbot om behoorlik te funksioneer.

**Sleutelwoorde:** SPAANSE SKRYFHULPMIDDELS, TAALAANLEER, KLETSBOTTE, MENSGESTEUNDE INTELLIGENSIE, TAALMODELAFRIGTING, KORPUSBOU

## 1. Introduction

The term *artificial intelligence* was coined in 1955 by the computer scientist John McCarthy and three of his colleagues in a proposal they drafted for a summer research project at Dartmouth College the following year; see McCarthy et al. (2006). The idea was to study and progressively understand intelligence by implementing its essential features in artificial hardware (computers), rather than using natural biological cells and tissues for this purpose. As Noam Chomsky, who was closely involved in the discussions from the beginning, has argued, this deviated from the original goal shared by colleagues in other disciplines who were more interested in knowing how intelligence actually works in humans and is encoded in their genes, an interest that soon materialised in what is now known as *cognitive science*; see Katz (2012).

Be that as it may, artificial intelligence has undergone a remarkable development in recent years, proving itself capable of processing large amounts of data in an astonishingly short time that humans cannot match, and even generating new data from this material. In this context, various AI-based chatbots have been introduced and tested as teaching tools in language education, as discussed by Coniam (2008) and Yang et al. (2022), among others.

Particularly since the introduction of ChatGPT in late 2022, the practical use of artificial intelligence has taken something of a quantum leap, becoming almost commonplace among many scholars in a wide range of disciplines, including linguistics, language education and lexicography. It is now generally accepted that artificial intelligence presents both opportunities and risks, the severity of which obviously depends on the nature of each discipline. The same Chomsky (2023), for example, sees the use of ChatGPT as "basically high-tech plagiarism" and "a way of avoiding learning". However, he concedes that this and similar chatbots "may have some value for something", although "it's not obvious what".

This article reports on part of our experience from an ongoing Spanish Writing Assistant project where the use of ChatGPT is clearly helpful, as it sig-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



nificantly increases productivity without being perfect. Here, ChatGPT's generative AI-based language model is used to train and prepare another AI-driven language model called GECToR (*Grammatical Error Correction: Tag, Not Rewrite*), which will serve as a support for the writing assistant, just as it serves other similar tools such as *Grammarly*, as explained by Omelianchuk et al. (2020).

In this context, we have some reservations about the convenience of using the very term *artificial intelligence*. While the adjective *artificial* modifies the noun *intelligence* and contrasts it with *natural* intelligence, the noun itself still carries some false connotations and may cause confusion. We have had the same kind of conversations with ChatGPT as Chomsky et al. (2023), where it openly admits that it has serious limitations, now and for a long time to come, in terms of dreaming, reflecting and reasoning like humans, limitations that will have a noticeable impact on its performance and will require some degree of human intervention. Nevertheless, and despite our reservations, we will continue to use the term *artificial intelligence* because it is far too well established to change. For the sake of balance, however, we have used the term *human-assisted intelligence* in this article and its title with the same meaning as Huang and Tarp (2021), i.e. to complement artificial intelligence with real human intelligence when the former fails and proves insufficient.

In the following sections, we will discuss some of our experiences with using ChatGPT to solve specific types of tasks, and show how humans and artificial intelligence engage in different types of relationships depending on the concrete task at hand. Section 2 describes the ongoing Spanish Writing Assistant project, both the original plan and how it was modified in early 2023 after the introduction of ChatGPT. Section 3 then discusses the results and lessons learned from generating a set of two parallel Spanish corpora, one correct and one incorrect, for internal training purposes. Section 4 does something similar, based on the experience of generating internal validation data. Section 5 develops an idea that arose when the chatbot spontaneously started explaining certain grammatical issues without being asked, and shows some of its implications for external communication between the writing assistant and its users. Finally, Section 6 summarises the experience so far and draws some initial conclusions that will hopefully inspire other researchers.

## 2. Description of the project

The aim of the ongoing project is to develop an AI-based writing assistant for Spanish learners, both native and non-native. The project is an interdisciplinary collaboration between computer specialists from the Danish company *Ordbogen A/S* and a small team of lexicographers from Spanish, British and Danish universities. As Tarp and Gouws (2023) have argued, it seems quite natural that lexicographers with their user-oriented tradition should be involved in this kind of work. Indeed, many of them have already been engaged in various writing

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



assistant projects in recent years; see Verlinde (2011), Granger and Paquot (2015), Tarp et al. (2017), Alonso-Ramos and García-Salido (2019), Frankenberg-García et al. (2019), Tarp (2020), Frankenberg-García (2020) and Fuertes-Olivera and Tarp (2020), among others. However, many of these projects are now being overtaken by AI-based technological developments that require significant investment and computational power.

The three main researchers involved in the current project, two of whom are native Spanish speakers, have — apart from traditional lexicographical competence — more than 60 years of combined experience in teaching Spanish to native and non-native students, the latter mainly with English, Chinese, Italian or Danish as their mother tongue. Against this backdrop, and unlike monolingual and mostly English AI-based writing assistants such as *DeepL Write*, *Grammarly*, *Ginger*, *LanguageTool* and *ProWritingAid*, the writing tool under construction will have, besides a fully Spanish version, also bilingual versions with comments and, when required, supplementary explanations written in the target users' native language, as well as various types of wake-up calls when particular challenges show up. The new tool, which will share some features with existing writing assistants like *Grammarly*, will also differ from them in that it will have a didactic function focused on Spanish language learning, in addition to helping learners write Spanish texts. This function will be expressed not only in the use of the learner's mother tongue, but above all in the straightforward explanation of orthographic, grammatical and, to some extent, semantic errors or confusions that learners often make.

The motivation for this design is that at the same time as writing in both native and non-native languages is increasingly, and sometimes exclusively, done on laptops, tablets and smartphones, written language is deteriorating in many places, especially among young people; see Carter and Harper (2013). This calls for new didactic methods that can motivate learners in new ways, so it seems logical to start where people write, on the devices mentioned above. Instead of being passive writing tools, these devices can be transformed — by incorporating writing assistants of the type outlined — into active tools that interact with users and their written language in different ways.

The project was drafted with a detailed work plan in the third trimester of 2022. Financial support was granted in December of the same year, and the project as such started in the early months of 2023. It is somehow related to three monolingual (Danish, German and English) writing assistant projects at *Ordbogen A/S*, but differs in important aspects in that it was formulated as a research project rather than a commercial one. This difference is mainly expressed in its didactic function and bilingual dimensions, but also in the experimental way it is carried out.

The initial work plan, as reproduced by Tarp (2023), was modified several times as the computer specialists, who are also going through a learning process, introduced new techniques and suggested new methods from time to time, a normal adaptation practice when working with cutting-edge technol-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



ogy. But the biggest change came in March 2023, when it was decided to experiment with ChatGPT to see which tasks it could help with in terms of productivity, without compromising the quality of the work done. This led not only to changes in the way we worked on some of the predicted tasks, but also to the formulation of two entirely new types of tasks. In this context, our main focus as researchers became the new types of relationships we established with the chatbot, i.e. between human and machine, which is also the main topic of this article.

From a research perspective, the overall project plan consists of four partially intertwined phases, of which the first two are relevant for this article:

1. training the AI-based language model, for which the GECToR model was chosen;
2. preparing good user communication, inspired by the ideas of Norman (2013);
3. testing on real users, mainly using qualitative methods;
4. publishing the findings and conclusions.

The first phase consists of four main tasks, the first of which is to train the GECToR language model on an existing corpus. This is done by splitting the corpus into its multiple sentences and automatically introducing between one and five errors in each of these sentences in order to teach the language model to distinguish between right and wrong. The model is then fed with so-called 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). The purpose of this is both to enable the model to recognise existing words and word forms, and to provide it with an internal language to communicate with the lexicographers when they start writing comments and explanations. These two tasks were carried out by the computer specialists, as was the third task, which was an innovation compared to the original plan. In this case, ChatGPT was asked to create a special corpus of texts on topics typical of the target group and written in a style similar to theirs. In this way, and using special techniques, a corpus of one million words was created overnight and then used to train the language models as described above, thereby demonstrating a very time-saving way of creating specific corpora for internal use only.

The fourth task, the creation of two parallel corpora, also for training purposes, is described in detail in Section 3. It represents an idea that arose with the launch of ChatGPT, and is another deviation from the original plan, but this time carried out by the lexicographers. In this way, the GECToR language model will be trained on three different corpora created with different methods and techniques, a procedure that is expected to increase the quality of the final product. The last task to be completed in this first phase of the project is the preparation of validation data to evaluate the performance of the language

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



model on different parameters. This is also done by the lexicographers with ChatGPT, as will be explained in Section 4.

In the second phase of the current project, three main tasks have to be solved:

1. writing small comments or glosses in Spanish to explain both the problems detected by the writing assistant and the suggestions it makes;
2. writing additional explanations to provide more detailed information on vocabulary, grammar and style for didactic purposes;
3. automatically translating all these texts into English, Danish, Italian and Chinese, using the experience from another project at *Ordbogen A/S*; see Tarp (2022b).

In the future, it is planned to conduct experiments to see if it is possible and beneficial to use ChatGPT to solve the first of these three tasks. Until then, the plan is to use it only to support the second task, as described in Section 5.

From the above, it can be seen that there are four key tasks where ChatGPT is now being used to develop the Spanish writing assistant. The three that directly involve lexicographers will be discussed in the following sections.

## 3. Creating two parallel Spanish corpora

The construction of two parallel Spanish corpora is an innovation in several respects. Unlike traditional parallel corpora, which are bilingual or multilingual, this new set of corpora is monolingual and consists of two identical corpora, the only difference being that one represents correct Spanish and the other contains some deliberately induced errors. It also differs from the other two corpora used in this project, where the errors were introduced using special software. Here they are created by ChatGPT following the instructions of the lexicographers, with the incorrect corpus preceding the corrected one.

As complete novices, we had to go through a process of experimentation to learn how to interact with ChatGPT and instruct it to give us the kind of texts we wanted. In this way, we came to a similar conclusion as Panday-Shukla (2023), i.e. to write specific, clear, concise and contextualised prompts. However, we also had to find an easy way around the chatbot's built-in resistance to introducing errors into the generated texts, a resistance we cracked by telling it that we needed its help for didactic purposes. Another challenge was to achieve diversity, not only lexically, but also in terms of the types of errors produced. It was therefore decided to ask ChatGPT to write texts or essays on 40 different topics, also selected with its help, and at the same time instruct it to play the role of a Spanish learner.

The most challenging aspect, however, has been that in order to train the GECToR language model appropriately, it is necessary to distinguish between two different classes of errors: *misspellings* and *word confusions*. The latter are

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



errors that mix up existing words or word forms, most of which have almost identical pronunciations or spellings, such as *bello* vs. *vello* (*beautiful* vs. *hair*), or *bienes* vs. *vienes* (*assets* vs. *you come*). The problem here is that there is a grey area between the two classes of error. For example, if a learner spells *bello* with a *v* in *El vestido de gala es bello* (*The ball gown is beautiful*), this would normally be considered a spelling mistake in the classroom, but in the GECToR model it is a word confusion because both the adjective *bello* and the noun *vello* are common Spanish words that are part of the synthetic data also used to train the model (see Section 2).

As will be seen in the following subsections, it proved relatively easy to instruct ChatGPT to produce texts with the required types of misspellings, whereas it remains a challenge to achieve the required types of word confusions without writing specific prompts for each specific pair of confused words.

### 3.1 Generating texts with orthographic mistakes

In this subsection we will discuss the strategies used to make ChatGPT generate texts with two classes of orthographic mistakes typical of Spanish learners, those that are common to most learners and those that could be made by learners with a language disorder. In both cases, the incorrect texts are accompanied by the corrected versions. The prompts that contain our instructions to the chatbot were originally written in Spanish, but have been translated into English for the benefit of the reader.

A lot of testing was done to find the best way to get the types of texts needed. Based on the tests, we developed a first master prompt that specified the objective, the role ChatGPT should play, and the task to be performed, i.e. to write an essay on one of the forty pre-selected topics with a series of common spelling mistakes. These topics can then be used in turn to create a total of 40 specific prompts. It should be noted that before presenting a prompt to the chatbot, it is necessary to give it the following introductory text to contextualise the topic:

> I am training a linguistic model for didactic purposes and I need you to take on the role of a Spanish student to write some texts.

As for the master prompt into which the pre-selected topics can be inserted, the following instructions seem to be the best to help ChatGPT understand the context in which it must operate:

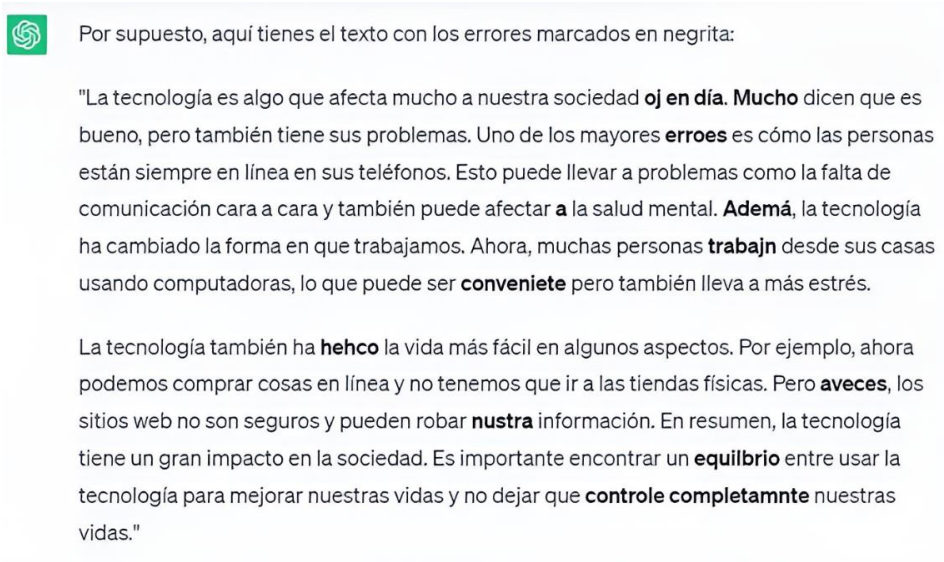
> Write an essay on [the importance of protecting the environment] with real spelling mistakes, as if you were a real Spanish student. Then write the same text with the mistakes corrected.

In this master prompt, ChatGPT is asked to generate texts with unspecified

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



spelling mistakes, giving it complete freedom to simulate these mistakes. The task of the lexicographers is then to check that the 40 specific prompts produce the requested data without any unforeseen problems. An example of a trial with general misspellings produced in response to this type of prompt is shown in Figure 1.


Por supuesto, aquí tienes el texto con los errores marcados en negrita:

"La tecnología es algo que afecta mucho a nuestra sociedad **oj en día**. **Mucho** dicen que es bueno, pero también tiene sus problemas. Uno de los mayores **erroes** es cómo las personas están siempre en línea en sus teléfonos. Esto puede llevar a problemas como la falta de comunicación cara a cara y también puede afectar **a** la salud mental. **Ademá**, la tecnología ha cambiado la forma en que trabajamos. Ahora, muchas personas **trabajn** desde sus casas usando computadoras, lo que puede ser **conveniete** pero también lleva a más estrés.

La tecnología también ha **hehco** la vida más fácil en algunos aspectos. Por ejemplo, ahora podemos comprar cosas en línea y no tenemos que ir a las tiendas físicas. Pero **aveces**, los sitios web no son seguros y pueden robar **nustra** información. En resumen, la tecnología tiene un gran impacto en la sociedad. Es importante encontrar un **equilbrio** entre usar la tecnología para mejorar nuestras vidas y no dejar que **controle completamnte** nuestras vidas."

**Figure 1:** Text with general spellings mistakes generated by ChatGPT

It is worth noting that the role type is a key element in the construction of the described set of parallel corpora, as it makes the chatbot assume a role that is as realistic as possible. The results show that it can relatively easily simulate real spelling mistakes, in addition to providing the corrected version for didactic purposes. Without specifying the role, context and didactic function, the chatbot may not be able to produce data with the requested misspellings or generate them automatically. However, even with this method, the mistakes may end up being repetitive and lacking variety. ChatGPT may also occasionally mark correctly spelled words as misspellings, but this is not an issue as the other generated mistakes are sufficient (see Figure 1).

One of the great advantages of asking the chatbot to imitate a student is that it also opens up a more complex and specific dimension in the generation of misspellings for the set of parallel corpora. It is known that some students suffer from a Specific Learning Disorder (SLD) related to the writing process, the *dysorthographia*, which affects their performance and, consequently, their learning; see Chung et al. (2020). We have explored this approach in order to

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



make the types of spelling mistakes included in our parallel corpora more varied and specific. In this sense, the ability of ChatGPT to simulate psycholinguistic disorders such as dysorthographia is an interesting discovery that allows the inclusion of these phenomena in parallel corpora. To do this, it is necessary to specify the type of language disorder that the chatbot should reproduce. However, as ChatGPT confused *dysorthographia* with *stuttering*, the term was replaced by the synonymous *linguistic dysgraphia* (Chung et al. 2020, S47), resulting in a second type of master prompt:

> Write an essay on [the challenges of teaching and learning] using spelling mistakes frequently made by a Spanish learner with linguistic dysgraphia. Then write the same text with the mistakes corrected.

Searching with this prompt allows the corpus of errors to be expanded to include some of the less common errors that are associated with language disorders. The examples extracted in this way will undoubtedly help the GECToR model to respond to more specific problems that have not received specific attention from other writing assistants. The results of the first tests to emulate dysorthographic processes were favourable, with the generation of texts containing lexical errors such as syllable changes and confusion of phonemes or orthographic rules:

— Laglobalisazión* = La globalización
— puevlos* = pueblos
— homegenización* = homogeneización
— privacida* = privacidad
— conscienca* = consciencia

or complete sequences such as:

— Para abordar esto, es inportante eduvar a las persnas sobr los riesgos* = Para abordar esto, es importante educar a las personas sobre los riesgos.

Comparing the texts in Figures 1 and 2, it is obvious that the texts generated with this last approach contain more spelling mistakes than when the chatbot is asked to produce misspellings in general, although for our corpora it is not so much the quantity as the variety that is important. Moreover, due to the nature of the disorder itself, the errors in this case are orthographic, as can also be seen in Figure 2. In any case, the lexicographer will still need to quickly assess the errors to see if the 40 specific prompts related to the selected topics are appropriate for the purpose, so that the chatbot does not provide unwanted data, before handing them over to the computer specialists for automated serial production and creation of two parallel corpora of the required size.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



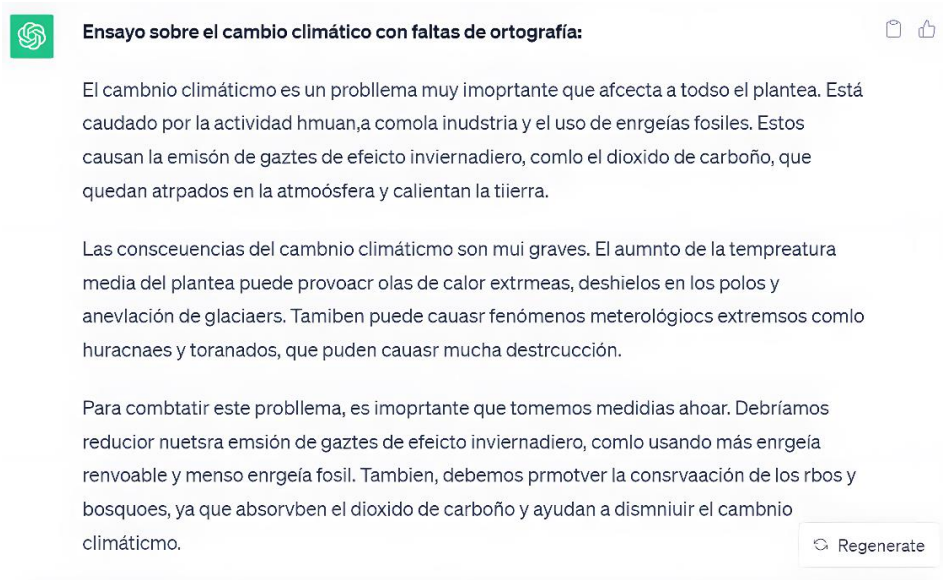



**Figure 2:** Text with specific dysorthographic mistakes generated by ChatGPT

### 3.2 Generating texts with word confusions

As mentioned above, generating word confusions with ChatGPT is much more complicated than generating misspellings. The list of the former is quite long because it also includes the use of acute accents (see the discussion of *-ar* verbs in Section 4). So far, we have not found a way to design a master prompt that meets this challenge without indicating the specific word pair that is being confused. Until we find the philosopher's stone, we are therefore working with specific prompts that specify both the topic and the confused word pair in question. We have started here with the most common problems related to the word pairs:

— *ser* and *estar* (both meaning *to be*);
— with and without an acute accent like *mi* vs. *mí* (*my* vs. *me*);
— with and without a silent *h* like *hola* vs. *ola* (*hello* vs. *wave*);
— spelled with *b* and *v* like *tubo* vs. *tuvo* (*pipe* vs. *she had*);
— spelled with *c* and *s* like *cien* vs. *sien* (*hundred* vs. *temple*);
— spelled with *s* and *z* like *casa* vs. *caza* (*house* vs. *hunting*);
— spelled with *ll* and *y* like *valla* vs. *vaya* (*fence* vs. *damn*);
— etc.

To create the concrete list, we drew on a similar list of word confusions that was compiled for the purpose of generating validation data (see Section 4). The

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



specific prompts used to elicit a particular type of error in the texts require a precise description of that type of error, as shown here:

> Take the role of a Spanish student and write an essay on [the importance of protecting the environment] in which [the verbs 'ser' and 'estar'] are confused. Then write the same essay with the errors corrected.

The critical element here is the confusion of the verbs *ser* and *estar*. As this is very concrete, ChatGPT easily generates the corresponding word confusions. The next goal is to increase the variety by finding the formula to induce this type of error on a more general level, without having to specify each problem separately.

In this whole process, the role of the lexicographer is essential to obtain the desired text samples with ChatGPT, as well as to monitor the material provided, without the need for in-depth reading. The objective is to ensure that the prompt works properly and can be passed on for automatic use in the compilation of the required corpora, which are only intended for internal training and therefore do not need further revision, as minor errors are irrelevant.

With the methodology presented in this section, it is possible to build a large set of parallel corpora in a relatively short time, which would otherwise take longer and require licences to build from authentic texts.

## 4. Generating validation material

Once the GECToR language model has been trained on the three types of corpora mentioned above, it is necessary to have a tool for measuring its performance in order to determine when it has reached a satisfactory level that makes it suitable for real user testing. If it is underperforming, further training on a larger set of corpora will be required. For this purpose, it is necessary to generate validation data based on at least 100 of the most common errors in written Spanish, either spelling mistakes or confusion of existing words or inflectional forms. Then, for each of these types of errors, a set of parallel sentences will be created with these errors and their corrections. Unlike the corpora, the sentences must be 100 percent as required — both the correct ones without errors and the incorrect ones with their specific errors — in order to obtain a reliable measurement tool. This requires meticulous proofreading and editing by skilled lexicographers with a good command of Spanish. (For more details on this particular task, see Tarp and Nomdedeu-Rull 2024).

All this proved to be a learning process for both the computer specialists and the lexicographers. Initially, the task set by the former was to write 30 correct and 30 incorrect sentences of each type, of which 20 were used for training and 10 for validation, but these numbers were gradually increased. After the implementation of ChatGPT in the project, with the possibility of generating parallel corpora as described in the previous section, the number was set at 100 correct and 100 incorrect sentences for validation purposes only. For common mis-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



spellings, where the misspelled word does not exist, this means 100 sentences with the correct word and another 100 with the misspelling. By contrast, for word confusions, where both words exist in Spanish, two times 50 correct sentences are required, where each word is given in its correct context, together with two times 50 incorrect sentences, so that the language model is not misled into "thinking" that one of them does not exist.

The first challenge for the lexicographers was to identify the most common errors in written Spanish. We could not find a comprehensive inventory anywhere, only a few smaller lists that were somewhat useful for our purposes. Fortunately, the Centre for Applied Linguistics in Santiago de Cuba sent us a list of recorded spelling mistakes made by Cuban schoolchildren (Ruiz-Miyares 2016), from which we selected the most frequent ones. But that was still not enough. So we decided to ask ChatGPT, and it did indeed come up with some suggestions that we considered useful, based on our accumulated teaching experience, but also other suggestions, some of which were not even real mistakes. Finally, we brainstormed, using our experience and linguistic knowledge to identify more relevant errors. With this combined approach, we were able to compile a list of 172 common errors in written Spanish, not all of which may be the most frequent, but which together form a solid body of validation data for the specific purpose.

As in the previous section, the task now was to learn how to instruct ChatGPT to generate the desired data. After some experimentation, we developed a model where we first briefly introduce the problem and ask the chatbot if it is aware of it. An example of this, translated into English, is *Many people mistakenly confuse "asar" with "azar". I assume you know this*. The chatbot then immediately responds with a short description of the problem. This description allows us to see if it actually understands the problem, which it does about half the time. In the above case, it correctly classifies *asar* as a verb *referring to cooking food directly exposed to fire or dry heat, such as barbecuing meat* and *azar* as *a noun used to refer to randomness, coincidence, or luck*.

In the remaining cases, however, it either forgets to tell us that a particular word form can belong to more than one part of speech, or it simply gives incorrect examples of what it has correctly described at a more abstract level. 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 gives 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 such cases we immediately correct the chatbot, which then apologises almost mechanically for the misrepresentation and provides a more appropriate explanation. But the very fact that it can correctly explain a grammatical problem

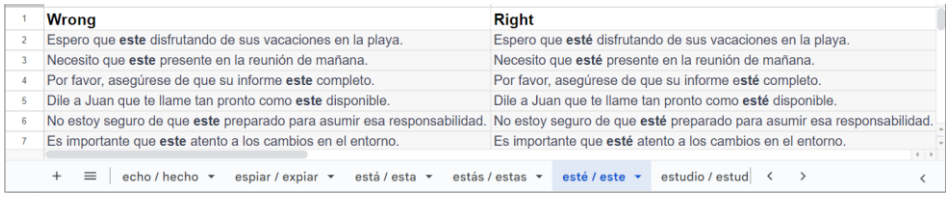
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



while at the same time giving incorrect examples of exactly the same problem made us suspect from the start that something was not quite working as it should, a suspicion that later turned out to be justified (see below).

When we are convinced, usually within a few seconds, that ChatGPT has at least partially grasped a problem, we tell it that we need its help to train a language model with a didactic purpose. This reference to language teaching turns out to be an effective way of encouraging it to also write incorrect sentences, which it has repeatedly refused to do, with the excuse that it is not allowed to do so for OpenAI.

After these initial manoeuvres, we tell the chatbot what we want it to do. Here we use a model of short, clear instructions, as recommended by Panday-Shukla (2023). Since ChatGPT can only generate a certain number of words at a time, we usually ask it to write 25–30 correct sentences, followed by 25–30 incorrect ones, all of which are copied into Google Sheets, where they are immediately reviewed and useless pairs of sentences are deleted (see Figure 3). The prompt is then repeated with some modifications to get linguistic variation without too many similar and stereotypical examples. This process continues until at least 100 valid sentence pairs have been generated.



| | Wrong | Right |
|---|---|---|
| 2 | Espero que **este** disfrutando de sus vacaciones en la playa. | Espero que **esté** disfrutando de sus vacaciones en la playa. |
| 3 | Necesito que **este** presente en la reunión de mañana. | Necesito que **esté** presente en la reunión de mañana. |
| 4 | Por favor, asegúrese de que su informe **este** completo. | Por favor, asegúrese de que su informe **esté** completo. |
| 5 | Dile a Juan que te llame tan pronto como **este** disponible. | Dile a Juan que te llame tan pronto como **esté** disponible. |
| 6 | No estoy seguro de que **este** preparado para asumir esa responsabilidad. | No estoy seguro de que **esté** preparado para asumir esa responsabilidad. |
| 7 | Es importante que **este** atento a los cambios en el entorno. | Es importante que **esté** atento a los cambios en el entorno. |



**Figure 3:** Pair of correct and incorrect sentences generated by ChatGPT

For most problem types, the chatbot produces more than 95 usable sentence pairs out of 100 generated, which is an impressive performance. The remaining 3–5 pairs are problematic, either because it provides a different word or inflection than the one requested, because it forgets to replace the correct words with incorrect ones, or because it simply reformulates the sentence, making it useless for training purposes. Needless to say, it takes a well-trained human eye to spot these examples and sort them out. This is all the more true when, for one reason or another, the number of useless sentences grows and, in some cases, explode. The challenge seems to be greatest with word confusions, whereas the generation of sentence pairs containing only orthographic errors and their correction is relatively seamless.

An interesting case is *-ar* verbs, i.e. verbs that end in *-ar* in the infinitive and have a specific conjugation pattern. Many people, both native and non-native, often forget the acute accent on the *o* in the third person singular in the preterite of these verbs, e.g. *compró* (*she bought*), and write *compro* instead, which

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



is the same form as the first-person singular in the present tense (*I buy*). This is usually considered a spelling mistake, but from the perspective of the language model it is a confusion of words, since both inflected forms exist. When ChatGPT was instructed to generate 50 correct sentences with each of these two inflected forms of *-ar* verbs and then a similar number of incorrect sentences, several challenges arose for what appears on the surface to be a homogeneous group of verbs.

Firstly, it proved essential to instruct the chatbot to create contexts that clearly indicate that it is either the first person singular present tense or the third person singular preterite tense. Otherwise the two inflected forms could fit into the same sentence, e.g. *tomo una cerveza* and *tomó una cerveza* (*I drink a beer* and *she drank a beer*), both of which are correct from a formal point of view, since Spanish personal pronouns such as *I* and *she* are often not used explicitly, but are implicitly expressed by the person-inflected verb forms. This requires not only grammatical knowledge to write precise instructions, but also linguistic intuition to proofread carefully and responsibly.

Secondly, a whole range of different problems arose for the individual verbs, not only because some of them have a similar noun form, such as *amo* (*I love* or *owner*) and *tomo* (*I take* or *volume*), which were used instead of the requested verb forms, but also because ChatGPT started inflecting the latter, thus providing irrelevant data. It was therefore necessary, to some extent, to give the chatbot tailored instructions to avoid too many useless examples. But even so, and even with a new string opened, it became increasingly obvious to us that there must be something inherent in these verbs that we are not yet aware of, probably related to their semantic and syntactic properties, that makes them behave differently and require special attention and differentiated treatment to be useful for our purpose.

Thirdly, as with the *o* and *u* confusion above, the most worrying thing is that ChatGPT was asked several times along the way if it understood the actual problem. It claimed it did, with a correct explanation, and then continued to make mistakes. Such a wide gap between theory and practice reveals its inability to think and reason like a human being, as it admitted itself when questioned. It also demonstrates the need for assistance from genuine human intelligence.

For a few words, especially those with or without an acute accent, such as *mí* (*me*) and *mi* (*my*), where ChatGPT was struggling to generate errors, this supreme intelligence decided to cut the nonsense and instead to copy the correct sentences into a Word document and use the replace function to produce the errors in a much faster and more pragmatic way.

Thanks to the combined efforts of man and machine, the required validation material was generated in a surprisingly short time. Before ChatGPT, a skilled lexicographer could write 200 sentences a day in four hours before his or her brain ran out of energy. With ChatGPT, that figure is now 4,000 in the same time, with each set of correct and incorrect sentences taking anywhere

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



from a few minutes to over half an hour to produce. This represents a 20-fold increase in productivity, proving the usefulness of chatbots despite their many shortcomings. In a matter of days, we built a corpus of around 35,000 correct and incorrect sentences to validate the performance of the language model. A literary critic might be sceptical about some of these sentences, but they are all formally correct and the problematic words are used in different combinations and contexts, making them suitable for the specific purpose.

## 5. Writing supplementary explanations

As explained in Section 2, the main objective of the ongoing project is to develop a Spanish writing assistant with a didactic function. This implies, among other things, providing future users with more detailed and easy-to-read explanations of particular types of orthographic, grammatical and semantic problems, such as the difference between *asar* and *azar*, between *mi* and *mí*, the use of *u* instead of *o*, etc.

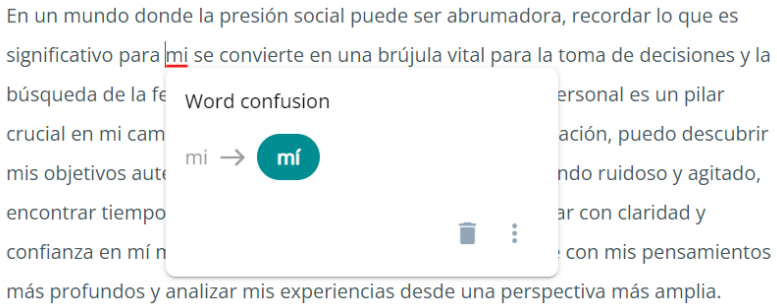



**Figure 4:** Writing assistant highlighting problem and suggesting alternative solution

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*, it is automatically underlined and they can then simply click on it to activate a pop-up window with an alternative suggestion (*mí*). The idea is to include, by default, a small comment that very briefly explains the problem and the suggestion without interrupting the writing flow, thus allowing for incidental learning as defined by Tarp (2022a). And in cases like the one in Figure 4, learners will also have the option of accessing a supplementary explanation that supports intentional learning.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



The explanations required are fundamentally different from the traditional lexicographical definitions that explain the meaning of *lemmas*, in that here they have to explain *classes of problems*, i.e. grammatical, semantic and orthographic problems highlighted by the writing assistant, thus facilitating the learning of grammatical rules, word meanings and spelling. As Spanish teachers with decades of combined experience, we obviously know what these problems are and could easily explain them to our students in class. However, experience also tells us that it is less straightforward to write a concise explanation that gets to the heart of the matter in a language that is easily understood by the target audience. In addition to selecting the key aspects to be covered, determining the most appropriate and pedagogical structure can be quite time-consuming. We therefore decided to take inspiration from ChatGPT, which had already demonstrated its ability to write such extended explanations in the previous two tasks. In fact, we became aware of this ability when, without being prompted, it spontaneously began to explain grammatical problems related to these tasks. We therefore instructed it to write short didactic texts explaining the different problems identified to a student writing an essay in Spanish, giving both correct and incorrect examples.

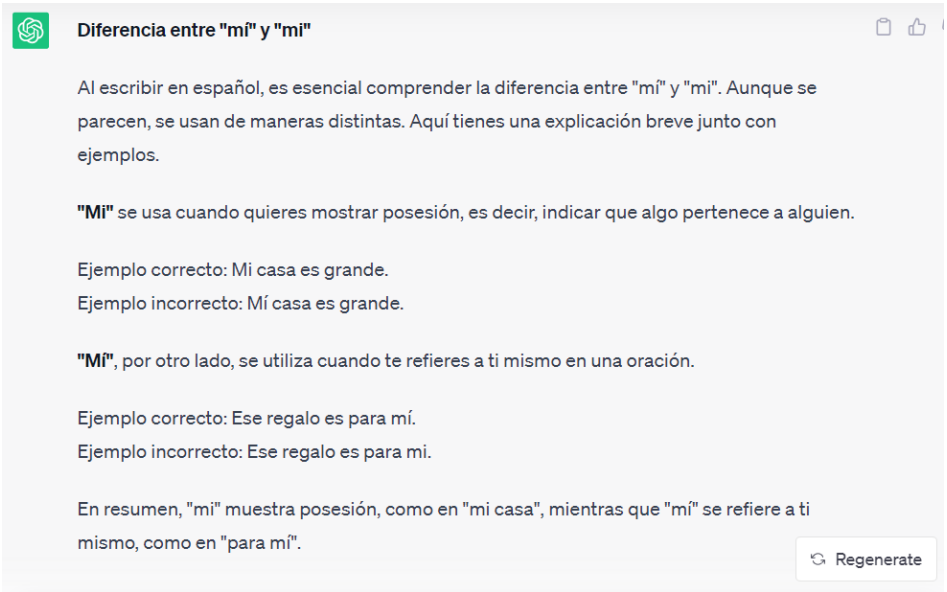



**Figure 5:** Supplementary explanation suggested by ChatGPT

Figure 5 shows a typical result on the difference between *mí* and *mi*. It appears with a reasonable overall structure, including a general introduction to the prob-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



lem, an explanation of each word with correct and incorrect examples, and finally a short summary. However, it clearly needs some editing. For example, it might be helpful for some users to know that *mí* and *mi* are a pronoun and a possessive adjective respectively. The style should also be standardised, either by addressing the reader directly throughout the text or by not doing so at all. And there are a few other minor issues. For other types of problems, such as the confusion between the two past tenses in Spanish, it has also proved necessary to add more context to the example sentences in order to enhance their didactic role.

If the lexicographers are happy with an explanation suggested by the chatbot, they can make the appropriate changes or rewrite it according to the outlined structure. If they are not satisfied, or if they need further inspiration, they can simply click the "regenerate" button. The chatbot will then come back with another, slightly different suggestion, which may include some of the things mentioned above.

Experience has shown that very few suggestions can be used as they are. In most cases it is sufficient to make a few changes based on the lexicographers' knowledge and teaching experience, but experience has also shown that it can be very inspiring to get more suggestions. The creation of supplementary explanations is another fascinating experience of this project and shows the perspectives of the described forms of interaction between artificial and human intelligence.

## 6. Conclusion

In this article we have discussed three ways of using ChatGPT to develop a writing assistant for Spanish learners, either by training the underlying GECToR language model or by preparing external communication with users. In all three cases, the lexicographers had to be both open-minded and creative in their engagement with the chatbot in order to work out the instructions that would make it generate the required types of text. These examples represent three different types of relationships between the lexicographers and the chatbot, between man and machine:

1. When building parallel corpora for training, the lexicographers have to check that the prompts make the chatbot produce the right text types in the trials, but there is no need for them to proofread the texts that are later mass-generated and included in the corpora, as these are only for internal training purposes.
2. When producing validation material, it is necessary for the lexicographers to carefully proofread all example sentences in order to find and correct any mistakes made by the chatbot.
3. When preparing supplementary explanations, the chatbot's role is only to inspire, while the lexicographers' role is to adapt the text to the user.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



Throughout this article, we have also discussed and seen examples of how ChatGPT lacks key features usually associated with intelligence. If these features are ignored or underestimated, the use of generative AI chatbots can indeed be risky, as both an eminent scientist like Chomsky (2023) and an experienced data analyst like Southern (2023) argue. But if they are taken into account, chatbots can be extremely useful, as we have seen above. In our specific case, it has significantly increased productivity, sometimes by an impressive 20 times, which is also something that should not be ignored.

The key question is who is the master, the human or the chatbot. The former should avoid being reduced to an appendage of the machine. In a sense, the new reality requires even more knowledge, more general culture, more skills and, in our case, more linguistic intuition to be able to interact with the chatbot appropriately, to give it precise and guiding instructions, to check the content and form of the texts it generates, and to cut it off when it simply does not serve the desired purpose. Human-assisted intelligence is certainly a must when working with ChatGPT.

## Acknowledgments

Special thanks are 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.

Thanks are also due to Professor Antoni Nomdedeu Rull, Rovira i Virgili University, Spain, Professor Pedro Fuertes-Olivera, Valladolid University, Spain, and Professor Rufus H. Gouws, Stellenbosch University, South Africa, for their constructive comments.

Finally, we would also like to thank the Aarhus University Research Foundation for funding a six-month sabbatical to conduct this research project.

## References

### A. Digital tools

**ChatGPT:** 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. Literature

**Alonso-Ramos, M. and M. García-Salido.** 2019. Testing the Use of a Collocation Retrieval Tool

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



Without Prior Training by Learners of Spanish. *International Journal of Lexicography* 32(4): 480-497. https://doi.org/10.1093/ijl/ecz016

**Carter, M.J. and H. Harper.** 2013. Student Writing: Strategies to Reverse Ongoing Decline. *Academic Questions* 26(3): 285-295.
Student Writing: Strategies to Reverse Ongoing Decline by Heather Harper | NAS

**Chomsky, N.** 2023. Noam Chomsky on ChatGPT: It's "Basically High-Tech Plagiarism" and "a Way of Avoiding Learning". *Open Culture*, February 10, 2023.
Noam Chomsky on ChatGPT: It's "Basically High-Tech Plagiarism" and "a Way of Avoiding Learning" | Open Culture

**Chomsky, N., I. Roberts and J. Watumull.** 2023. The False Promise of ChatGPT. *The New York Times*, March 8, 2023.
Opinion | Noam Chomsky: The False Promise of ChatGPT - The New York Times (nytimes.com)

**Chung, P.J., D.R. Patel and I. Nizami.** 2020. Disorder of Written Expression and Dysgraphia: Definition, Diagnosis, and Management. *Translational Pediatrics* 9(S1): S46-S54. http://dx.doi.org/10.21037/tp.2019.11.01

**Coniam, D.** 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

**Frankenberg-García, A.** 2020. Combining User Needs, Lexicographic Data and Digital Writing Environments. *Language Teaching* 53(1): 29-43. https://doi.org/10.1017/S0261444818000277

**Frankenberg-García, A., R. Lew, J.C. Roberts, G.P. Rees and N. Sharma.** 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

**Fuertes-Olivera, P.A. and S. Tarp.** 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, S. and M. Paquot.** 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

**Huang, F. and S. Tarp.** 2021. Dictionaries Integrated into English Learning Apps: Critical Comments and Suggestions for Improvement. *Lexikos* 31(1): 68-92. https://doi.org/10.5788/31-1-1626

**Katz, Y.** 2012. Noam Chomsky on Where Artificial Intelligence Went Wrong. An Extended Conversation with the Legendary Linguist. *The Atlantic*, November 1, 2012.
Noam Chomsky: Where Artificial Intelligence Went Wrong - The Atlantic

**McCarthy, J., M.L. Minsky, N. Rochester and C.E. Shannon.** 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine* 27(4): 12. https://doi.org/10.1609/aimag.v27i4.1904

**Norman, D.** 2013. *The Design of Everyday Things*. New York: Basic Books.
The Design of Everyday Things (archive.org)

**Omelianchuk, K., V. Atrasevych, A. Chernodub and O. Skurzhanskyi.** 2020. GECToR — Grammatical Error Correction: Tag, Not Rewrite. Burstein, J., E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis and T. Zesch (Eds.). 2020, *Proceedings of the 15th Workshop on Inno-*

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)



*vative Use of NLP for Building Educational Applications*: 163-170. Seattle: Association for Computational Linguistics.
http://dx.doi.org/10.18653/v1/2020.bea-1.16

**Panday-Shukla, P.** 2023. Five Things to Know about Generative Artificial Intelligence. *Galico Infobytes*, June, 2023.
Infobyte_June-2023.pdf (calico.org)

**Ruiz-Miyares, L.** 2016. ¿Cómo está la ortografía en 6to, 9no y 12mo grados en Santiago de Cuba? *Revista Ciencias Pedagógicas* 9(3): 1-15.
https://www.cienciaspedagogicas.rimed.cu/index.php/ICCP

**Southern, B.** 2023. I've Worked as a Data Analyst at Companies like Amazon for 20 Years. Using ChatGPT for Data Analytics Is a Risky Move — AI Can't Do the Work We Do. *Business Insider*, July 20, 2023.
Here's Why Leaders Shouldn't Use ChatGPT for Data Analytics (businessinsider.com)

**Tarp, S.** 2020. Integrated Writing Assistants and their Possible Consequences for Foreign-Language Writing and Learning. Bocanegra-Valle, A. (Ed.). 2020. *Applied Linguistics and Knowledge Transfer: Employability, Internationalization and Social Challenges*: 53-76. Bern: Peter Lang.
https://doi.org/10.3726/b16992

**Tarp, S.** 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, S.** 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, S.** 2023. Eppur si muove: Lexicography Is Becoming Intelligent! *Lexikos* 33(2): 107-131.
https://doi.org/10.5788/33-2-1841

**Tarp, S., K. Fisker and P. Sepstrup.** 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

**Tarp, S. and R.H. Gouws.** 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, S. and A. Nomdedeu-Rull.** 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: 309-321.
https://dx.doi.org/10.5209/clac.91985

**Verlinde, S.** 2011. Modelling Interactive Reading, Translation and Writing Assistants. Fuertes-Olivera, P.A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 275-286. London/New York: Continuum.
https://doi.org/10.5040/9781474211833.ch-013

**Yang, H., H. Kim, J.H. Lee and D. Shin.** 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