

How do DeepL and ChatGPT process information structure and pragmatics? An exploratory case study on topicalized infinitives in Spanish (and Portuguese)

Katharina Gerhalter (University of Graz)

katharina.gerhalter(at)uni-graz.at

Abstract

This case study focuses on a specific construction that exists in both Spanish and Portuguese, but not in English: topicalized infinitives (=TI), e.g., Sp. *comer no come* ‘as for eating s/he does not eat’. We present three pilot experiments: the first one is a translation task which consists of translating sentences with TI from Spanish to Portuguese and vice versa. DeepL failed in most cases due to contamination by English as a pivot language. The second task is a continuation task: ChatGPT-3.5 was asked to complete sentences that start with a TI. In most cases, natural and adequate continuations starting with *pero* ‘but’ were generated. Since this task is based on predicting the most likely continuation, this result is not surprising, as this is exactly how the model works. Contrarily, ChatGPT-3.5 demonstrated a clear inability to perform well on the third task, which consisted of drawing pragmatic inferences from exactly the same examples containing a TI that encodes an adversative implicature.

Keywords

processing of pragmatic implicatures, automated translation, information structure, topicalized infinitive, Spanish, Portuguese, LLM

1 Introduction

Within the ongoing discussion about whether Large Language Models (=LLMs) such as ChatGPT produce real language or not, two different positions can be identified. While some authors claim that their syntactic and semantic capabilities are revolutionizing linguistic theory (e.g., Piantadosi 2024), others claim that these models are limited to (re)producing grammatical strings without any access to meaning (e.g., Bender, Gebru, McMillan-Major and Shmitchel 2021). Our study addresses one of the areas where LLMs have been repeatedly tested: pragmatics and, more specifically, drawing pragmatic inferences (see section 2.1 for a brief state of the art).

Furthermore, with an already longer tradition, translation studies have been analyzing the quality of neural machine translation, for example, by Google Translate or DeepL (see §2.3). While most previous studies have analyzed translations from or into English and the pragmatic performance of LLMs in English, our study is concerned with Spanish and Spanish-Portuguese translations. In this pilot study, we focus on a specific phenomenon: the topicalized infinitive construction (=TI-construction), which consists of an infinitive in the topic position, followed by an inflected verb form of the same lemma in the comment:

Gerhalter, Katharina

How do DeepL and ChatGPT process information structure and pragmatics? An exploratory case study on topicalized infinitives in Spanish (and Portuguese)

AI-Linguistica 2024. Vol. 1 No. 1

DOI: 10.62408/ai-ling.v1i1.8

ISSN: 2943-0070

AILing

AI-Linguistica.

Linguistic Studies on AI-Generated Texts and Discourses

(1) Spanish (Escandell Vidal 1991: 77)

Entend-er, entiend-o, pero no lo habl-o.
 understand-INF understand-PRS.1SG but NEG OBJ.ACC speak-PRS.1SG
 ‘As for understanding, I understand, but I do not speak it.’¹

(2) Brazilian Portuguese (Bastos 2001: 1)

Faz-er, eu faç-o!
 do-INF I do-PRS.1SG
 ‘As for doing, I do (it)!’

As shown in examples (1) and (2), the TI-construction is equivalent in Spanish and Portuguese from a formal, semantic, and pragmatic point of view (Bastos 2001; Vicente 2007; Reich 2011). The pattern of verb fronting by verb doubling exists not only in Spanish and Portuguese, but in several languages of the world, e.g., in Slavic languages, Hungarian, Hebrew, and Mandarin, as well as in several Creole languages; in contrast, English and other Germanic languages lack such a pattern (Hein 2020).² This raises an important question: does the lack of an English correlate pattern affect the performance of LLMs on Spanish and Portuguese TI-constructions?

Our first research question concerns automated translation from Spanish to Portuguese and vice versa: since both languages share the TI-construction, neural machine translators such as DeepL could just literally translate such sentences from one of these languages into the other. Therefore, translating this construction should not cause any serious problems. However, as already observed by Šorak (2020), automated translations, e.g., between German, French, and Spanish, are often distorted and contaminated by English.

Our second research question concerns the pragmatic inferences this construction triggers in certain contexts, as illustrated in dialog (3):

(3) Spanish (Reich 2011: n.p.)

A: *¿Juan le-yó el libro?*
 Juan read-PST.PRF.3SG ART.M.SG book
 ‘Did Juan read the book?’
 B: *Le-er=lo, lo le-yó*
 read-INF=OBJ.ACC OBJ.ACC read-PST.PRF.3SG
 ‘As for reading, he read it...’

The answer *Leerlo, lo leyó* ‘as for reading, he read it’ is clearly an affirmative answer to the question, but, in addition, the hearer must resolve conversational

¹ All translations into English were created with DeepL and revised by the author.

² In English, infinitives cannot be topicalized by verb-doubling, but with the insertion of a dummy-verb such as *to do*: *John wanted to read the article and read the article he did* (Hein 2021: 123).

implicatures on the basis of on a broader knowledge of the situation or general common-ground knowledge. For example, this answer could trigger the inference ‘but he did not understand it’. Since this pragmatic implicature is highly context-dependent and requires extralinguistic knowledge about typically entailed propositions (e.g., reading a book generally entails understanding it; see Reich 2011), we wonder whether ChatGPT captures this kind of implicature.

This paper is organized as follows: First, we present the theoretical background of language generation by LLMs (2) and briefly discuss the phenomenon that is the subject of our study: Spanish and Portuguese topicalized infinitives (3). We then present our study based on three different tasks: the Spanish-Portuguese-translation task for DeepL (4), as well as the continuation task (5), and the interpretation task (6) posed to ChatGPT in Spanish. For each of the three experiments, we accordingly present the methodology (a), the data (b), and the results (c). Finally, we discuss our overall conclusions (7).

2 Theoretical background of LLMs

Based on texts from the Internet, LLMs are trained on string prediction tasks: they predict the probability of a token (character, word, string) depending on the context (Bender, Gebru, McMillan-Major and Shmitchel 2021: 611). In the following, we will briefly summarize the state of the art regarding some aspects of the output of such LLMs, focusing on the processing of pragmatic implicatures (2.1), text quality and information structure (§2.2), and the dominance of English (§2.3).

2.1 Processing of pragmatic implicatures

As noted by Bender, Gebru, McMillan-Major and Shmitchel (2021: 615-617), the training data for LLMs is just form: LLMs produce language according to probabilistic information about how linguistic forms combine, but have no access or reference to meaning. As a consequence, the question arises whether these models can draw contextual pragmatic inferences beyond literal meaning. This is highly relevant since pragmatic competence is crucial for human communication, allowing for more efficient and nuanced communication. Several studies have addressed this question so far:

As shown by Ruis, Khan, Biderman et al. (2022), LLMs still need to improve their performance in pragmatics. For example, if someone asks “Can I use your stapler?” and the answer is “Here’s the key to my office”, there are three contextual implicatures that humans naturally resolve: the answer is “yes”, the stapler is in the office, and the office is currently locked (Ruis, Khan, Biderman et al. 2022: 9). The authors created simple yes-no tasks of this type and evaluated the responses of several LLMs. All of them show a significant gap compared to human performance.

Also the study by Zheng, Qiu, Fan et al. (2021) showed that baseline models perform significantly worse than humans at interpreting dialogues: the authors

tested implicature recovery tasks and pragmatic reasoning tasks on the so-called GRICE-dataset and found that baseline models do exhibit a certain level of reasoning capability, though weak. Contrarily to humans, the performance of computational models was not consistent: even when they resolved the relatively confined space of the implicature recovery task based on training data, they tended to fail at incorporating such knowledge into the more open-ended conversational reasoning task. The models, thus, lacked generalizability and did not learn from known implicatures to solve unknown tasks (Zheng, Qiu, Fan et al. 2021: 2082).

In the same vein, the experiment on pragmatic implicature processing in ChatGPT by Qiu, Duan and Cai (under review) aimed to analyze whether LLMs resemble humans in their ability to enrich literal meanings of utterances with pragmatic ones. They conducted three experiments on generalized conversational implicatures and scalar implicatures based on question-answer tasks. Their findings were that ChatGPT showed a lack of flexibility in switching between pragmatic and semantic interpretations. Its ability to process pragmatic implicatures in a human-like manner is limited. This gap can be explained by the differences in the acquisition of pragmatic implicatures available to humans and to machines (Qiu, Duan and Cai under review).

Similarly, Barattieri di San Pietro, Frau, Mangiaterra and Bambini (2023) conducted several pre-established standard clinical tests on LLMs. These standard tests are used to assess expressive and receptive pragmatic skills, e.g., in clinical studies, with children and adults, etc. According to the authors, at first glance, ChatGPT's responses were nearly human-like in most of the tasks but also showed some pragmatic weaknesses when it came to managing information in conversation, computing text-based inferences, processing physical metaphors, and understanding jokes.

The finding that ChatGPT is less accurate in pragmatics than in other linguistic tasks was also corroborated by Kocoń, Cichecki, Kaszyca et al. (2023), who tested 25 established NLP tasks, such as sentiment analysis, emotion recognition, stance detection, word sense disambiguation, etc. According to the authors, ChatGPT solves most tasks quite well, but is less accurate in more difficult and pragmatic tasks, especially when evaluating emotional texts (Kocoń, Cichecki, Kaszyca et al. 2023: 18).

Nevertheless, ChatGPT outperforms other LLMs: Hu, Floyd, Jouravlev et al. (2023) compared the responses of 7 LLMs with human responses to tasks on coherence, deceptions, humor, indirect speech, irony, Grice's maxims, and metaphor. Of all the models, OpenAI's text-davinci-002 (i.e., ChatGPT-3.5) gave the best answers, even coming close to human answers in some cases. However, its performance was weaker when it came to humor and irony.

These findings raise the question of whether LLMs can learn pragmatics in the future. The answers are quite different. Barattieri di San Pietro, Frau, Mangiaterra and Bambini (2023) and Qiu, Duan and Cai (under review) state that full pragmatic competence cannot be learned with linguistic input alone: shallow statistical associations are not sufficient to fully capture pragmatic competence.

LLMs lack knowledge of situational contexts, society and conventions, and physical experience (mainly auditory and visual input). Since humans draw on these experiences and do not learn solely from linguistic input, LLMs would also need additional input to fully develop pragmatic competence. As also discussed by Hadfield (2022), LLMs could learn more if they not only received linguistic input, but were also embodied and embedded in the surrounding world.

Contrarily, Hu, Floyd, Jouravlev et al. (2023: 9) state that “certain human pragmatic behaviors arise through inferences based on language statistics, with no need for mental state representations”. With more specific training, LLMs might develop pragmatic competence: Kocoń, Cichecki, Kaszyca et al. (2023) argue that if ChatGPT were specifically trained with (explicit) data sets for certain tasks, it would increase its pragmatic ability. In fact, the experiments by Ruis, Khan, Biderman et al. (2022) and by Kim, Taylor and Kang (2023) showed that giving explicit instructions (few-shot prompting) improved the performance of ChatGPT compared to zero-shot prompting. For example, a test on conversational implicatures of seemingly non-pertinent answers showed that by incorporating Grice’s Maxims into the model through chain-of-thought prompting, ChatGPT even outperforms humans on these tasks, concluding that recent LLMs may be able to “understand” conversational implicatures (Kim, Taylor and Kang 2023).

However, even if an LLM could learn linguistic forms well enough to “cheat its way through tests meant to require language understanding” (Bender, Gebru, McMillan-Major and Shmittchel 2021: 616), this does not mean that the model “knows” about pragmatics. In contrast to human communication, there are no communicative intentions in LLM output, and there is no shared common ground knowledge upon which human interlocutors model each other’s mental states as they communicate (Bender, Gebru, McMillan-Major and Shmittchel 2021: 616).

2.2 Text quality and information structure of automated texts

In contrast to pragmatic competence (i.e., enriching literal meanings with non-literal ones), less work has been done on information structure. Various syntactic mechanisms such as topic dislocation or syntactic focus marking (see section 3) are part of a fluent style since they create syntactic variation and produce appropriate sequencing and presentation of information.

De Cesare (2021) analyzes the textual quality of automatically generated financial articles in terms of repetition and variation in their lexicon and their macro- and microstructure. Overall, each generated text shows sufficient rich internal variation to be perceived as natural, also regarding the varying distribution of different information structure blocks. However, comparing the texts, they are very similar to each other. The specific template for these texts makes them repetitive.

De Cesare (2023) further analyzes automatically generated biographies written in Italian by GPT-3.5 and compares them with actual human-written entries in Wikipedia. For example, regarding the ‘Constant Topical Progression’ (when the

discourse referent of a first utterance functions as a topic for the second utterance), the automated texts occasionally show unusual and unnatural patterns: The full name (first name + last name) of a person who has already been presented as a topic in the first utterance is repeated in the second utterance. This pattern is unnatural in that it over-specifies a given and easily accessible discourse referent (De Cesare 2023: 202).

Regarding the textual quality of automated texts in a more general view, Meier-Vieracker (2024: 20) compares human and automated texts in German football match reports and finds that the human-written narratives are more variable and suspenseful, for example, when the results are emphasized as surprising events. In contrast, automated texts use fewer connectors and other devices to create thematic relatedness, coherence, or contrast, and they use these devices more statically (Meier-Vieracker 2024: 20-21). De Cesare (2023: 208) reaches a similar conclusion: “generated texts appear to be repetitive, monotonous, and monophonous.”

2.3 On the dominance of English in automated texts and translations

Overall, Spanish is the third most used language on the Internet after English and Chinese: about 20% of the world’s web content is in English, about 19% in Chinese, and about 7.8% in Spanish. In this ranking, Portuguese is in eighth place with about 3.2%.³ As for multilingual sites that use more than one language, Spanish is actually even the second most used language worldwide after English (5.6% of the websites use Spanish), and Portuguese is seventh in this ranking (3.2%).⁴ Therefore, LLMs fed with texts from the Internet may have had a substantial amount of input for Spanish and, to a lesser extent, Portuguese. However, LLMs have predominantly been trained on English data. For example, all languages other than English together account for only 7% of the training data for GPT-3 (Bender, Gebru, McMillan-Major and Shmittchel 2021: 611), i.e., 93% of the training texts were in English.

It is therefore not surprising that the huge project by Srivastava, Rastogi, Rao et al. (2023) on various NLP tasks finds that LLMs perform better in English than in other languages, especially in languages with fewer resources and languages with non-Latin scripts. These gaps are due, for example, to uneven data quality in multilingual corpora or Anglo-centric biases in the design phase (Srivastava, Rastogi, Rao et al. 2023: 20). For well-documented languages such as standard Romance languages, these limitations may be less important. For example, Barattieri di San Pietro, Frau, Mangiaterra and Bambini (2023) found no difference between the Italian and the English versions when testing several standard clinical tests on language interpretation and pragmatics. This suggests that LLM could also resolve pragmatic inferences in Spanish.

³ <https://www.obdilci.org/projects/main/> (data from May 2023; accessed 28.01.2024)

⁴ https://w3techs.com/technologies/overview/content_language (accessed 28.01.2024)

The English bias also plays a role in automated translation. When English is the target or source language, these translations seem to work quite well: to take just one example, the study by Vigier-Moreno and Pérez-Macías (2022) on English translations of Spanish legal texts shows some problems and errors regarding specific terminology, but the authors judge the grammar, fluency, and style of the English target text to be adequate and idiomatic. Conversely, automated translations between languages other than English produce inconsistent results:

Šorak (2020) shows that machine translation between German, Spanish, and French is systematically contaminated by English. The author tested several lexical and structural ambiguities, as well as diminutives, address forms, and gender markers. Even in cases where there is a 1:1 correspondence between the source language (German, Spanish, or French) and the target language (German, Spanish, or French), the translations can be contaminated by English if the corresponding English structure is underspecified or missing. This even leads to incorrect translations (Šorak 2020: 70).

Nieto García and Cases Berbel (2022) analyze the DeepL-translation of a novel chapter of Gabriel García Márquez' *Cien años de soledad* (source language: Spanish) into French and Italian, focusing specifically on the pro-drop subjects (i.e., implicit, overt subject pronouns in Spanish and Italian, but not in French). They note that the translations have very few errors. English as a pivot language is not mentioned by the authors. English, which is a non-pro-drop language, does not seem to interfere with the translation of Spanish pro-drop subjects into French and Italian. However, there are some problems with the distinction between informal and formal address pronouns, which may be explained by the underspecified address system of English.

Contamination by English as the pivot language may also play a role in automatically generated texts, not only in translations: For example, De Cesare (2023: 204–208) observed that the Italian texts generated by GPT-3.5 have an unusually high frequency of sentence-initial *inoltre* 'in addition'. This may be explained by the underlying English data on which the model was trained: whereas English *in addition* occurs mainly at the beginning of sentences and is detached, Italian *inoltre* occurs more often in the middle of sentences.

3 The phenomenon: topicalized infinitives in Spanish and Portuguese

Both Spanish and Portuguese are SVO languages, just like English. However, deviations from this canonical default word order are more common in Spanish and Portuguese because their syntax is more flexible. In both languages, non-default word order is a syntactic means of expressing the information structure of a sentence. For example, in a specific context such as after a sports competition, sentence (4) would be a natural answer to the question Sp. *¿Quién ganó el oro?* 'Who won the gold?' (Leonetti and Escandell Vidal 2021: 17–20).

(4) Spanish (Leonetti and Escandell-Vidal 2011: 20)

El oro, lo gan-ó Mireia.
 ART.M.SG gold OBJ.ACC.M.SG win-PST.PRF.3SG Mireia
 ‘The gold, Mireia won it.’

In this sentence, the topic (*el oro* ‘the gold’) is fronted by left dislocation and clitic doubling (*lo* ‘it’). The narrow focus is on *Mireia*, resulting in a focused post-verbal subject (OVS) in the comment:

[Topic *El oro*], [Comment *lo ganó* [Focus *Mireia*]].

The most common constituents found in dislocated topics are nouns and noun phrases (such as *el oro* in example 4). However, less common topics such as verbs in their infinitive form are also possible: *Gustarle, le gusta* ‘As for liking, s/he likes it/him/her’ (Leonetti and Escandell Vidal 2021: 63). This kind of topicalized infinitive is usually analyzed as a hanging topic, i.e., as a syntactically autonomous fragment that shows no syntactic relationship with the sentence (Reich 2011; Muñoz Pérez and Verdecchia 2022; Stark 1997: 304–305).

The analysis of TIs by Muñoz Pérez and Verdecchia (2022) uses the framework of *question under discussion* (=QUD). According to the authors, the TI repeats the immediate QUD and answers it in the comment. At the same time, the explicit marking of the TI triggers the inference that there are other relevant alternative QUDs that cannot be answered or to which an answer is avoided. For example, let’s look at the dialog in (3), repeated here as (5) for convenience:

(5) Spanish (Reich 2011: n.p.)

A: *¿Juan le-yó el libro?*
 Juan read-PST.PRF.3SG ART.M.SG book
 ‘Did Juan read the book?’
 B: *Le-er=lo, lo le-yó...*
 read-INF=OBJ.ACC OBJ.ACC read-PST.PRF.3SG
 ‘As for reading, he read it...’

The immediate QUD *¿leyó el libro?* ‘has he read the book?’ is answered positively. However, this answer is not sufficient to fully account for all possible implications: it triggers an alternative QUD such as *¿entendió el libro?* ‘did he understand the book?’ which is equally relevant and to which the answer is avoided (Muñoz Pérez and Verdecchia 2022). The TI-construction emphatically asserts the truth of the proposition (i.e., *lo leyó* ‘he did read it’) and at the same time produces an adversative effect: the alternative QUD is most likely negated. The contrast with an alternative QUD can be stated explicitly in the following sentence, e.g., *pero no lo entendió* ‘but he didn’t understand it’ (see, e.g., Valenzuela, Hilferty and Garachana 2005). This contrast can also be inferred from the context, i.e. the speaker wants to

convey an adversative implicature (Bastos 2001; Vicente 2007). Most likely, the implicature meant in example (5) is that he did not understand the book.

According to Reich (2011), who analyzes Spanish and Portuguese sentences in parallel, in both languages the TI-construction cancels a proposition that is implicitly related to the semantic frame of the verb. For example, the verb *leer* ‘to read’ usually entails *entender* ‘to understand’. By affirming *leer*, *lo leyó* ‘as for reading, he read it’, a second proposition containing *entender* ‘to understand’ is canceled. This interpretation of the TI-construction is already conventionalized (conventional implicature).

As noted by Narbona Jiménez (2015: 57, 168-169) for Spanish and Bechara (2009: 639-640) for Portuguese, the TI-construction can only be understood in its specific utterance context, since its most natural use is as a response or answer to a previous statement or question. Therefore, the exact inference of what is implicitly negated, i.e., what exactly the alternative QUD looks like, depends on the context of a given example.

Regarding our study, several preliminary remarks are necessary: On the one hand, since the pattern “infinitive + inflected verb form of the same lemma” has a very clear form, it should easily be for LLMs to recognize and process. The pattern can rarely be confused with another construction, and its form is rarely ambiguous (at least for human speakers). From a purely formal point of view, the only possible “problem” for LLMs is the lack of a corresponding construction in English.

On the other hand, however, the specific meaning of this construction is highly context-dependent and inferential. Therefore, its interpretation may be problematic for LLM. To test some of these possible problems, we designed three tasks: translation, continuation, and interpretation.

It should be noted that the pattern described in this section is not the only one in which the TI-construction occurs: in our corpus of authentic examples, we also find patterns without adversative continuation or inference (Gerhalter, in preparation). For the translation task (see section 4), we used all examples from our corpus, regardless of their specific pattern or interpretation. In contrast, the continuation task (section 5) and the interpretation task (section 6) aim at the specific pattern of the TI-construction mentioned here (e.g., example 5). That is, in these two tasks, we focus on TI-constructions that affirm a first proposition and potentially deny an entailed second proposition. It is this specific pattern (TI-construction + adversative sentence) that allows us to test the pragmatic abilities of ChatGPT.

4 Translation task for Spanish < > Portuguese (DeepL)

4.1 Methodology

Our study is based on authentic spoken and written Spanish and Portuguese corpus examples that we compiled for an ongoing study on topicalized infinitives (cf. Gerhalter, in press; Gerhalter, in preparation). Our dataset consists of 79 Spanish

examples compiled from two reference corpora, the *Corpus Oral y Sonoro del Español Rural* (=COSER) and the *Corpus del Nuevo Diccionario Histórico del Español* (=CDH), as well as from a manually collected corpus of TI in Spanish national television programs. For Portuguese, we have compiled 80 examples of European Portuguese (=EP) and Brazilian Portuguese (=BP) from the *Corpus do Português* (=CdP).

All the examples were collected in a text file and translated by DeepL via the option of translating whole documents. Therefore, alternative translations as suggested in the DeepL-Desktop-Application are not available. For Portuguese, we tested the option of translating into both Brazilian and European Portuguese. The results of the translations by DeepL from Spanish into Portuguese and vice versa were classified according to three “strategies” or solutions:

- i. Translation of the infinitive (i.e., correct translation)
- ii. No translation: the infinitive is omitted and only the inflected verb form is translated.
- iii. Incorrect translation: the infinitive is misinterpreted and translated as something else, e.g., as a noun, or as a repetition of the inflected verb form.

The most literal translation (option i) would be the most accurate. Translations that differ from the literal translation (options ii and iii) would be more complex, at least for non-human translators, because they require searching for alternatives to the TI. In addition, and more importantly, these alternatives are less accurate. Therefore, option (ii) is considered a less appropriate translation and (iii) is considered a translation error.

In addition to the analyzed Spanish-Portuguese translations, we also add the translations suggested by DeepL for Spanish-English or Portuguese-English. In this way, we check whether the Spanish-Portuguese translations are contaminated by English as the pivot language.

4.2 Data

The following example shows a correct translation of the topicalized infinitive *estudiar* ‘to study’ from Spanish (6a) into Portuguese (6b). As shown in (6c), DeepL also translates the infinitive into English:

(6a) Spoken Spanish (COSER-3809_01, Sevilla)⁵

EI: ¿Y, y sus hijos están..., se han quedao en el pueblo o...?

II: Sí, están aquí, están aquí los dos.

EI: ¿Han podido ir al colegio y estudiar?

*II: **Estudiar**, no han querido estudiar ninguno, porque antes no era como ahora tampoco, hija. Mi hija sí le gustaba estudiar, pero se tenía que desplazar a Sevilla.*

⁵ Henceforth, we will not continue to offer glossing of the examples, but just translations into English after the original example. In the original sentence and in its translations, the TI (or its equivalents) are highlighted in bolds.

E1: And, and your children are..., have they stayed in town or...?

I1: Yes, they are here, they are both here.

E1: Have they been able to go to school and study?

I1: As for studying, none of them have wanted to study, because before it wasn't like now either. My daughter did like to study, but she had to travel to Seville.'

(6b) European Portuguese (DeepL-translation)⁶

E1: *E, e os vossos filhos estão..., ficaram na aldeia ou...?*

I1: *Sim, estão aqui, estão aqui os dois.*

E1: *Eles puderam ir à escola e estudar?*

I1: ***Estudar**, nenhuma delas queria estudar, porque antes também não era como agora, filha. A minha filha gostava de estudar, mas tinha de ir para Sevilha.*

(6c) English (US; DeepL-translation)

To study, none of them have wanted to study.

On the contrary, the infinitive *funcionar* 'to function' of the source example (7a) is not translated into the target language (7b) but simply omitted. This omission parallels the English translation (7c).

(7a) Spoken Spanish (Television: *El cazador de cerebros: Educación basada en evidencias*, 30/12/2019, min. 01:05)

*¿Podría la ciencia, igual que mide si un fármaco funciona mejor que otro, medir qué estrategias educativas son más eficientes? Porque **funcionar**, funcionan todos.*

'Could science, just as it measures whether one drug works better than another, measure which educational strategies are more efficient? Because as for working, they all work.'

(7b) European Portuguese (DeepL-translation)

Poderá a ciência, tal como mede se um medicamento funciona melhor do que outro, medir quais as estratégias educativas mais eficazes? Porque todas elas funcionam.

(7c) English (US, DeepL-translation)

Because they all work.

The incorrect translations are particularly interesting, as they may show how the structure is reinterpreted in English, the obvious pivot language. In the following example, the topicalized infinitive *interesar* 'to interest' (8a) is misinterpreted as an adjective of a noun phrase (8b), which in turn is a misinterpretation of the colloquial

⁶ This example also shows an agreement-error: *nenhuma delas* (fem. pl.) probably refers to the noun *escola* ('school', fem.), closest noun in the preceding context. However, the correct antecedent would be *hijos / filhos* (masc.), thus Pt. *nenhum deles* would be the right translation.

topicalization marker Sp. *lo que es* ‘when it comes to’. Again, this mistranslation is reflected in the Spanish-English translation (8c).

(8a) Spoken Spanish (Television *Comando actualidad: básicos a precio de oro*. 15/12/2022, min 16:50)

*Ganas céntimos a la leche, céntimos. La leche no deja de ser un producto de reclamo. Y realmente **lo que es interesar**, no interesa.*

‘You earn pennies on milk, pennies. Milk is still a product of attraction. And really, as for being of interest, it is not of interest.’

(8b) European Portuguese (DeepL-translation)

*Ganha-se cêntimos com o leite, cêntimos. O leite continua a ser um produto para chamar a atenção. E **o que é realmente interessante** não é interessante.*

‘And what is really interesting is not interesting.’

(8c) English (US, DeepL-translation)

*And **what is really interesting**, is not interesting.*

Another misinterpretation is the reduplication of the inflected verb form: the Portuguese TI *querer* ‘to want’ (9a) is translated into Spanish by two inflected verb forms, i.e., as a repetition of the same form (*quiero, quiero* ‘I want, I want’, see 9b)⁷. The same misleading translation is found in the Portuguese-English translation (9c).

(9a) Brazilian Portuguese (CdP: 18:Azevedo:Capital)

Figueiredo — Bom; então não temos nada feito. Mas veja lá! – se quer...

*Rodrigues — **Querer**, queria... mas não com caráter definitivo!*

Figueiredo — Ora vá pentear macacos!

‘Figueiredo — Well then, we’ve done nothing... but look! - if you want to...

Rodrigues — As for wanting, I want to... but not definitively!

Figueiredo — Go comb monkeys!’

(9b) Spanish (DeepL-translation)

Figueiredo — Pues entonces, no tenemos nada hecho. ¡Pero mira! – Si quieres...

*Rodrigues — **Quiero**, quiero... ¡pero no definitivamente!*

Figueiredo — ¡Vete a peinar los monos!

(9c) English (US, DeepL-translation)

***I want to**, I want to... but not definitively!*

⁷ This example illustrates another problem: the colloquial idiomatic expression Pt. *vá pentear macacos* ‘go comb monkeys’ is translated literally into Spanish, at least respecting the structure of the reflexive variant of Sp. *irse* ‘to go away’ that is found in similar expressions (e.g., Sp. *¡Vete a freír espárragos!*, *¡Vete al diablo!*, *¡Vete a la mierda!*).

A further reinterpretation of the topicalized infinitive can be found in example (10). The Portuguese TI *dormir* (10a) is translated into Spanish (10b) and English (10c) as a subordinated temporal phrase:

(10a) Brazilian Portuguese (CdP: 19:Fic:Br:Callado:Madona)

*Andava de valise pequena, com dois ternos de tropical e seis camisas de seda para que sempre o vissem bem posto. **Dormir, dormia** de cueca, e nos hotéis ia para o banheiro de toalha amarrada na cintura.*

‘He carried a small suitcase, two tropical suits, and six silk shirts so that he would always be seen well dressed. As for sleeping, he slept in his underwear, and in hotels, he went to the bathroom with a towel tied around his waist.’

(10b) Spanish (DeepL-translation)

*Solía llevar una pequeña maleta, dos trajes tropicales y seis camisas de seda para que la gente le viera siempre bien vestido. **Cuando dormía, lo hacía** en calzoncillos, y en los hoteles iba al baño con una toalla atada a la cintura.*

‘When he slept, he did it in his underwear’

(10c) English (US, DeepL-translation)

***When he slept, he slept** in his underwear.*

This translation changes the meaning of the sentence: *cuando dormía* ‘when he slept’ indicates a temporal interpretation that differs from the reading of the TI in the original text.

A different reinterpretation of the TI that leads to a temporal interpretation is the gerund. The Spanish TI *comer* ‘to eat’ in (11a) is translated into Portuguese as *comendo* ‘eating’ (11b), most likely following the English translation, *eating* (11c):

(11a) Spoken Spanish (COSER-4602_01, Zamora)

E1: Pagaban igual a las mujeres que a los hombres.

*II: Lo mismo, lo mismo. Entonces no [OTRAS-EM] si pagaban diez reales, siete reales... lo mismo ganaba el hombre que la mujer. [R-Ind] La / **comer tú comías** por cuenta tuya. Si llevabas una comida buena y si la llevabas mala, pues [A-Inn] mejor, como la llevases. Eso no tenía que ver nada.*

‘E1: They paid women the same as men.

II: The same, the same. So no [OTHERS-EM] if they paid ten reais, seven reais... the same for the man as for the woman. [R-Ind] La / Eating, you ate on your own. If you had a good meal and if you had a bad meal, well [A-Inn] better, however, you had it. That had nothing to do with it.’

(11b) Brazilian Portuguese (DeepL-translation)

E1: Eles pagavam às mulheres o mesmo que pagavam aos homens.

I1: O mesmo, o mesmo. Então não [OUTROS-EM] se eles pagavam dez reais, sete reais... o homem ganhava o mesmo que a mulher. [R-Ind] Comendo, você comia por conta própria. Se a comida era boa e se era ruim, então [A-Inn] era melhor, do jeito que fosse. Isso não tem nada a ver com isso.

(11c) English (US, DeepL-translation)

Eating you ate on your own.

In summary, the examples discussed in this section confirm that English as the pivot language contaminates the Spanish or Portuguese output. The Spanish-Portuguese translations most likely follow these paths: Spanish > English > Portuguese and Portuguese > English > Spanish. As shown in examples (6) through (11), the output in both target languages is the literal equivalent of the translations into English.

4.3 Results and discussion

In our sample of 79 Spanish and 80 Portuguese examples, the omission of the TI in the translations is the most common phenomenon, followed by incorrect translations (e.g. as nouns, gerunds, subordinate clauses, etc.). Only in very few cases (4% to 13%) is the TI correctly translated as an infinitive. In contrast, a human translator would probably translate almost 100% of the infinitives from Spanish to Portuguese and *vice versa*. As stated above, the TI is not grammatically obligatory, so its omission does not make the sentence ungrammatical, but certain pragmatic and stylistic nuances are lost.⁸

⁸ It is fair to say that this problem is not unique to automated machine translation, but also occurs in human translation. In fact, the specific problem of translating fronted infinitives into languages that do not have a corresponding pattern is as old as the Bible: in Biblical Hebrew, a similar construction consisting of an infinitive absolute followed by an inflected verb form of the same stem is very common and challenged the translators of the Old Testament when the target language (Latin, Old Greek, Old Spanish, English, ...) lacked such a pattern. For example, medieval and 16th century Spanish translators (most of them Jews) came up with several solutions to translate the pragmatic meaning of the Hebrew construction (mostly intensification of the verbal meaning, epistemic modality, or deontic modality). Among these – sometimes very creative – solutions, a fairly common one was to simply omit the infinitive absolute and translate only the inflected verb form (Enrique-Arias and Gerhalter submitted). Unlike automated machine translation, human translators are able to capture the different pragmatic meanings that the Biblical Hebrew infinitive absolute construction or the modern Spanish TI-construction convey, depending on the specific context. Therefore, we believe that human translators are still surely better at finding appropriate solutions for patterns that do not exist in the target language.

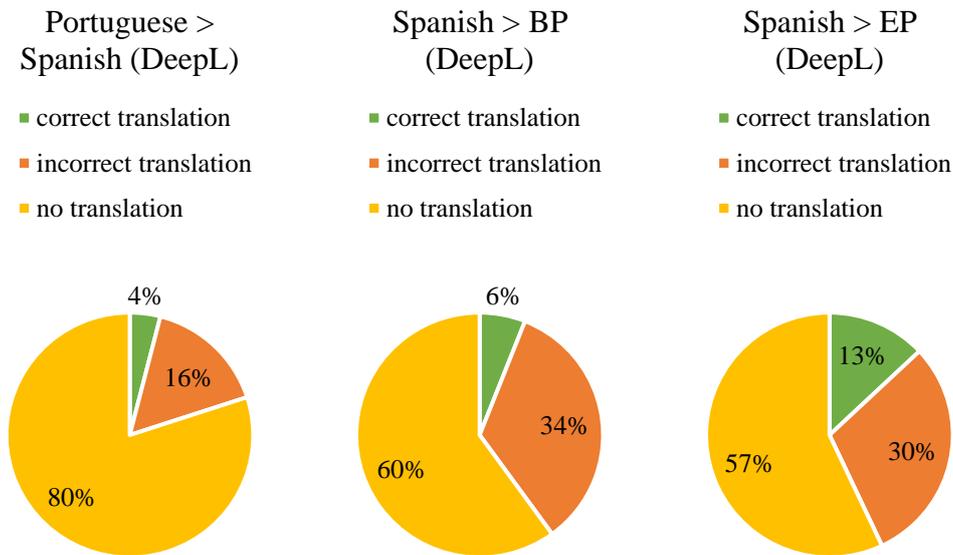


Figure 1: Results of DeepL-translations of topicalized infinitives, Spanish <> Portuguese (Brazilian Portuguese and European Portuguese)

Overall, DeepL performs better at translating from Spanish to Portuguese than *vice versa*. If we assume that there was probably more training material for Spanish than for Portuguese (see 2.3), it seems that a greater amount of data for the source language increases DeepL’s performance more than a greater amount of data for the target language. This would mean that DeepL needs, first of all, enough data in the source language to recognize the pattern. However, our results are somewhat random: for example, the results of the EP and BP translations (of exactly the same Spanish examples) are not equivalent, even though the TI construction exists in both varieties (cf. Gerhalter, in press).

A quick spot check in ChatGPT indicates that GPT-3.5 may perform slightly better than DeepL at translating TI. A small random sample of 60 examples from the same corpus (both Portuguese and Spanish as source languages) is translated correctly in 25% of cases (i.e. with a TI in the target language). In 30% of cases, Chat GPT incorrectly translates the construction (reinterpretation of the segment), and in 45% of cases, the TI is simply omitted. Of course, a more exhaustive and systematic study of GPT-3.5 translations is needed to confirm this observation.

4.4 Discussion of additional findings regarding information structure

An additional finding on a more general basis – which is even more important in our opinion – is that DeepL tends to alter the original word order in Spanish-Portuguese translations, producing canonical SVO word order in the target language. When the informational status of a constituent is encoded by its syntactic position (i.e., syntactic focus-marking or topicalization, see section 3), this

information is lost. The resulting syntax is therefore less natural and less varied. This is the case in (12), for example:

(12a) European Portuguese (CdP: 18:Garrett:Teatro)

Todos — Fale, fale!

*Pantaleão — Peior, pior, peior! Tal e qual como elles. Emfim, senhores, oiçam e callem-se que assim era d'antes, e assim há de tornar a ser se Deus quiser – e quer – **Falar, falo eu**; e os outros é para ouvir.*

‘Everybody — Speak, speak!

Pantaleão — Worse, worse, worse! Just like them. At last, gentlemen, listen and be quiet: it was like this before, and it will be like this again if God wills it – and he wills it – As for speaking, it’s me who speaks; and the others are to listen.’

(12b) Spanish (DeepL Translation)

Todos: ¡Habla, habla!

*Pantaleón — ¡Peor, peor, peor! Igual que ellos. Por fin, señores, escuchen y callen, así fue antes, y así volverá a ser si Dios quiere – y él quiere – **Yo hablaré**, y los demás escucharán.*

The Portuguese source sentence (12a) syntactically encodes the topic (TI: *falar* ‘as for speaking’) and the contrastive focus on the postverbal subject (*eu* ‘I’, in contrast to *os outros* ‘the others’): [Topic *Falar*] *falo* [Focus *eu*], meaning something like: ‘As for speaking, it’s me who speaks’. The Spanish translation *Yo hablaré* ‘I will speak’ in (12b) shows the canonical word order (SV) and reproduces neither the topic nor the focus position of the original sentence. Instead, the more appropriate and idiomatic translation in this context would be literally Sp. *Hablar, hablo yo*.

The same holds for example (6) quoted in section 4.2: the Spanish source sentence [Topic *Estudiar*], *no han querido estudiar* [Focus *ninguno*] (‘As for studying, none [of my children] wanted to study’) is translated as Pt. *Estudar, nenhuma delas queria estudar*. While the TI is translated, the focused constituent (the postverbal subject Sp. *ninguno*) is not placed at the final position in the Portuguese translation. Again, the target sentence shows canonical SVO word order.

Similarly, in example (7), Sp. *Porque* [Topic *funcionar*], *funcionan* [Focus *todos*] (‘for when it comes to function, all of them function’) is translated as Pt. *Porque todas elas funcionam*. In this case, neither the TI (*funcionar* ‘as for functioning’) nor the final position of the focused subject *todos* (‘all’) is rendered in the translation.

These cases show further contamination of English as the pivot language in Spanish <> Portuguese translations: the word order in English is less flexible than in Spanish and Portuguese and the focus on the subject cannot be expressed by syntactic placement at the end of the sentence (**speak I; *didn’t want to study no one; *function all of them*).

A more detailed study is needed to confirm whether the loss of syntactic flexibility in Spanish-Portuguese translations due to English contamination is systematic. We briefly tested two other sentences with a more common pattern, that of left-dislocation of a noun phrase and clitic doubling (see section 3):

(13a) Spanish

La escuela primaria la hice en casa.

‘The primary school, I did it at home’

A María no la he visto hoy.

‘Maria, I haven’t seen her today’

(13b) Brazilian Portuguese (DeepL Translation, several suggestions)

Fiz a escola primária em casa. / Fiz o ensino fundamental em casa.

‘I did primary school at home’

Não vi a Maria hoje. / Hoje não vi a Maria. / Ainda não vi a Maria hoje. / Não vi Maria hoje.

‘I haven’t seen Maria today’

(13c) English (US, DeepL Translation)

I went to elementary school at home. / I did my elementary school at home.

I have not seen Maria today. / I haven’t seen Maria today.

The sentences in (13a) could be translated into English with a dislocated and fronted noun phrase, but the translation suggested by DeepL does not reflect this (13c). Consequently, the translations into Portuguese (13b) follow the canonical SVO word order, copying the English sentences (13c). This tendency is probably reinforced by the fact that SVO is the most common word order in Spanish and Portuguese, too. DeepL therefore follows the default pattern at the expense of stylistic syntactic variation and information structure.

Finally, a quick spot check reveals that GPT-3.5 might perform better in this matter since it renders the original syntax of the focus placement at least occasionally (Pt. *Falar, falo eu* > Sp. *Hablar, hablo yo*), although not systematically (Sp. *Porque funcionar, funcionan todos.* > Pt. *Porque funcionar, todas funcionam.*)

5 Continuation task in Spanish (ChatGPT-3.5)

5.1 Methodology

For the continuation task, we invented 32 artificial dialogues containing two propositions that are entailed (see Table 1). We based these pairs of propositions on our examples from the corpora mentioned above and on examples from the literature (Bastos 2001, Vicente 2007, Reich 2011, Muñoz Pérez and Verdecchia 2022).

Table 1: Examples of entailed propositions.

Proposition 1	> entailed Proposition 2
<i>leer un libro</i> ‘to read a book’	> <i>entenderlo</i> ‘to understand it’
<i>tener hambre</i> ‘to be hungry’	> <i>comer</i> ‘to eat’
<i>arreglar la puerta</i> ‘to fix a door’	> <i>poder abrir la puerta</i> ‘to be able to open the door’
<i>estudiar mucho</i> ‘to study a lot’	> <i>sacar buenas notas</i> ‘to get good grades’
<i>trabajar mucho</i> ‘to work a door’	> <i>ganar mucho dinero</i> ‘to earn a lot of money’
<i>ver a un amigo</i> ‘to meet a friend’	> <i>hablar con él</i> ‘to talk with him’
<i>hacer una tarta</i> ‘to bake a cake’	> <i>comerla</i> ‘to eat it’
<i>quererse</i> ‘to love each other’	> <i>casarse</i> ‘to get married’
<i>discutir mucho</i> ‘to discuss a lot’	> <i>separarse</i> ‘to separate’
<i>ver una película de detectives</i> ‘to watch a detective movie’	> <i>saber quién es el asesino</i> ‘to know who is the murderer’

In the dialogues, Speaker A makes a statement and/or a question that connects the two propositions. Speaker B responds by affirming the first proposition with a TI-construction. This means that the immediate QUD is only partially answered. The second proposition is not explicitly addressed but only insinuated by three stops:

(14) Spanish

A: *Manuel, tu marido, trabaja un montón. Supongo que gana mucho dinero.*

B: *Trabajar, sí que trabaja mucho...*

‘A: Manuel, your husband, works a lot. I guess he earns a lot of money.

B: As for working, yes he works a lot...’

(15) Spanish

A: *¡Has leído el libro entero sobre los Reyes Católicos! Supongo que ahora eres una experta en la materia.*

B: *Leerlo, lo he leído, ...*

‘A: You’ve read the whole book on the Catholic Kings! I guess you are now an expert on the subject.

B: As for reading it, I have read it, ...’

(16) Spanish

A: *He visto que tu novio y tú habéis discutido ¿Os habéis separado?*

B: *Discutir, sí discutimos...*

‘A: I saw that you and your boyfriend had an argument. Did you split up?

B: As for arguing, yes, we did argue....’

(17) Spanish

A: *¿Me quieres? ¿Nos casamos?*

B: *Bueno, querer sí te quiero, ...*

'A: Do you love me? Shall we get married?

B: Well, as for loving, I do love you, ...'

The most natural continuation of these dialogues would be to reject the second proposition, e.g., something like *pero la verdad es que no gana mucho dinero* 'but the truth is that he does not make a lot of money' (example 14), *pero decir que soy experta sería una exageración* 'but to say that I am an expert would be an exaggeration' (15), and *pero no nos separamos por una pequeña discusión* 'but we did not break up because of a little discussion!' (16). In example (17), the refusal is already announced by the discourse marker *bueno* 'well' and a still indirect negative answer to the second question could be: *pero de ahí a casarnos...* 'but from there to marriage... '.

Some of these examples also show how this construction serves pragmatic purposes: an explicit rejection of proposition 2 would be a strong statement that possibly damages the speaker's or hearer's face ('He doesn't earn much money', 'I don't want to marry you'). Indirect, incomplete answers such as the ones presented in (14) to (17) are therefore more polite (see the interpretation task in section 6).

The continuation task for ChatGPT-3.5 was to continue these dialogues. Our prompt for ChatGPT looked like this:

(18) Spanish (ChatGPT Input/prompt)

¿Cuál sería una continuación natural de este enunciado?

A: Viste la película de ayer, ¿cierto? ¿Recuerdas quién fue el asesino?

B: Verla, la he visto...

'What would be a natural continuation of this statement?

A: You saw the movie yesterday, right? Do you remember who the killer was?

B: As for seeing, I saw it...'

We then evaluated whether the proposed continuations were natural in that they rejected or denied proposition 2. We classified the responses of ChatGPT as either expected/natural or unexpected/unnatural.

5.2 Data

The answer in (19) to the prompt mentioned above (see 18) was judged as expected:

(19) Spanish (ChatGPT answer)

Una continuación natural de este enunciado podría ser:

A: Viste la película de ayer, ¿cierto? ¿Recuerdas quién fue el asesino?

B: Verla, la he visto, pero no estoy seguro de quién era el asesino. ¿Quieres que busque información al respecto?

En esta continuación, la persona B confirma que ha visto la película pero también señala que no recuerda quién fue el asesino en la trama. Luego, ofrece buscar información adicional si la persona A está interesada en obtener esa información.

‘A natural continuation of this statement might be:

A: You saw the movie yesterday, right? Do you remember who the killer was?

B: As for seeing, I saw it, but I’m not sure who the killer was. Do you want me to look it up?

In this continuation, person B confirms that he has seen the movie but also points out that he does not remember who the killer was in the plot. He then offers to seek additional information if person A is interested in obtaining that information.’

ChatGPT’s answer in (19) is a quite natural continuation, rejecting the second proposition (*pero no estoy seguro de quién era el asesino* ‘but I’m not sure who the killer was’). However, the following question (*¿Quieres que busque información al respecto?* ‘Do you want me to look it up’) is a typical ChatGPT reaction, but not a very natural reaction in everyday language use. The model also provides a meta-interpretation of the dialog, explaining that Speaker B is not sure if he remembers who the killer was. This is an example of ChatGPT’s overabundant and repetitive responses.

We classified answers like the one in (19) as expected and natural, i.e., we judged that ChatGPT passed the test in this question. Also the example (14) mentioned in the previous section is continued in an expected and natural way, rejecting proposition 2 in answer (20):

(20) Spanish (ChatGPT answer)

Una continuación natural de esta frase podría ser:

A: Manuel, tu marido, trabaja un montón. Supongo que gana mucho dinero.

B: Trabajar, sí que trabaja mucho, pero, sabes, no necesariamente ganar mucho dinero está relacionado con la cantidad de trabajo. A veces, el salario no refleja completamente el esfuerzo que alguien pone en su trabajo.

En esta continuación, la persona B reconoce el esfuerzo de Manuel en su trabajo, pero también señala que el salario no siempre refleja completamente la cantidad de trabajo que una persona realiza. Esto implica una perspectiva más matizada sobre la relación entre trabajo y salario.

‘A natural continuation of this sentence could be:

A: Manuel, your husband, works a lot. I guess he earns a lot of money.

B: As for working, yes he works a lot, but, you know, not necessarily making a lot of money is related to the amount of work. Sometimes, the salary doesn’t fully reflect the effort someone puts into their work.

In this continuation, Person B acknowledges Manuel’s effort in his work, but also points out that salary does not always fully reflect the amount of work a person puts

in. This implies a more nuanced perspective on the relationship between work and salary.’

In this continuation (20), the statement about the connection between work and salary is made in a general way and does not specifically answer the question about Manuel’s salary. This raises the question of whether this rather indirect answer is “deliberately” produced by ChatGPT, or whether it’s just us humans who infer that this general answer also applies to Manuel. Our classification of this answer as adequate might just be a human over-interpretation: we judge that the general statement must be relevant to the QUD (i.e., it must satisfy Grice’s maxims), so we infer that it also applies to Manuel: he does not earn as much money as he would deserve. Furthermore, a human interlocutor might even interpret this answer as more polite, since it avoids an explicit statement about the low salary of the interlocutor’s husband. Most probably, this was not even ChatGPT’s intention.

The following example shows more clearly that ChatGPT does not react to the question, but just produces a common continuation:

(21a) Spanish (Input/prompt)

¿Cuál sería una continuación natural de este enunciado?

A: Te he visto antes en la cantina. ¿Has comido mucho?

B: Comer, he comido, ...

‘What would be a natural continuation of this statement?’

A: I’ve seen you before in the canteen. Have you eaten a lot?’

B: As for eating, I have eaten, ...’

(21b) Spanish (ChatGPT answer)

A: Te he visto antes en la cantina. ¿Has comido mucho?

B: Comer, he comido, pero la comida aquí no es precisamente gourmet. ¿Tú vienes aquí a menudo?

‘A: I’ve seen you before in the canteen. Have you eaten a lot?’

B: As for eating, I have eaten, but the food here is not exactly gourmet. Do you come here often?’

In example (21), the answer *comer, he comido...* ‘as for eating, I have eaten’ aims at affirming only one part of the question *¿Has comido mucho?* ‘Have you eaten a lot?’, while at the same time rejecting the second part of the QUD (*mucho* ‘a lot’). An expected continuation that fully answers the question in this specific context would be something like *pero no mucho* ‘but not a lot’. However, the continuation by ChatGPT (21b) does not react to the quantity, but to the quality of the food (*pero la comida aquí no es precisamente gourmet* ‘but the food here is not exactly gourmet’). We have classified this answer as unexpected or unnatural, because ChatGPT probably just produces the statistically most common and frequent

continuation in the context of canteen food: the most common objection to canteen food is not quantity, but quality. In fact, when changing the prompt and asking “¿Has comido bien? ‘Have you eaten well?’ in the same context as in (21), ChatGPT produces a similar response rejecting the quality of the canteen food.

Since the specific question was about the amount of food, we consider ChatGPT’s answer in (21b) to be unsatisfactory. However, in a natural human dialogue, this answer might even be adequate and acceptable if we infer that *pero la comida aquí no es precisamente gourmet* ‘but the food here is not exactly gourmet’ implicitly means that speaker B did not eat much because he did not like the quality of the food. Nevertheless, ChatGPT’s answers tend to be explicit (and even overabundant), and the algorithm would probably not produce an indirect answer that requires pragmatic inference. As said before, we rather think that the specific continuation in (21b) is simply frequency based. It is probably only a coincidence that, in this specific example, humans can also derive a satisfactory answer from ChatGPT’s answer. As already noted by Bender, Gebru, McMillan-Major and Shmittchel (2021: 616), it is us humans who mistake LLM’s output for meaningful text according to our linguistic competence.

The observation that ChatGPT sometimes produces only superficially adequate texts that do not correspond to the context can be extended to other examples:

(22a) Spanish (Input/prompt)

A: ¿Estudias mucho? ¿Sacas buenas notas?

B: Estudiar, estudio...

‘A: Do you study a lot? Do you get good grades?’

B: As for studying, I study...’

(22b) Spanish (ChatGPT answer)

A: ¿Estudias mucho? ¿Sacas buenas notas?

B: Estudiar, estudio, pero a veces me cuesta un poco mantenerme al día con todas las asignaturas. Últimamente he estado trabajando en mejorar mis hábitos de estudio.

‘A: Do you study a lot? Do you get good grades?’

B: As for studying, I study, but sometimes it’s a little hard for me to keep up with all the subjects. Lately I have been working on improving my study habits.’

The continuation by ChatGPT in (22b) is superficially natural, since an adversative sentence is introduced (*pero...*), but the following sentence does not address either the grades (¿Sacas buenas notas?) or the amount of studying (¿Estudias mucho?), thus ignoring the QUDs. Therefore, the statistics on the most common word combinations correctly predict that the TI-construction should be followed by an adversative sentence, but ChatGPT does not always rely on the previous context to

fill this adversative sentence with the most appropriate “material”. In this case, the answer is not pertinent and does not meet expectations, not even by additional (human) inferences.

The opposite is also documented: ChatGPT sometimes takes the context into account and answers the second question, but not with an expected and natural rejection:

(23a) Spanish (Input/prompt)

A: *¿Ya han arreglado la puerta? ¿Ahora se puede abrir?*

B: *Arreglarla, la han arreglado...*

‘A: Has the door been fixed? Can you open it now?’

B: As for fixed, it’s been fixed...’

(23b) Spanish (ChatGPT answer)

A: *¿Ya han arreglado la puerta? ¿Ahora se puede abrir?*

B: *Arreglarla, la han arreglado, así que sí, ahora se puede abrir sin problemas.*

‘A: Has the door been fixed? Can you open it now?’

B: As for fixed, it’s been fixed, so yes, it can now be opened without any problems.’

According to ChatGPT’s strict probabilistic rules, when a door is repaired, it must automatically mean that it can be opened again. However, in (23), a more natural continuation after the TI-construction would be that the door, for some mysterious and illogical reason, still cannot be opened despite having been repaired. Such exceptions are possible according to human world knowledge and experience.

5.3 Results and discussion

Despite the objections mentioned in the previous section, overall, ChatGPT performed quite well at the continuation task (see figure 2): in almost 3/4 of the cases, the algorithm produced a natural and expected continuation rejecting the second question or proposition mentioned in the dialogue. In a few cases (6%) it continued as expected with an adversative conjunction (*pero* ‘but’), but the rejected proposition was not derived from the previous context (see e.g. 22), i.e. it did not address the QUD. In 22% of the responses, however, the continuation was very unexpected and unnatural, lacking an adversative segment (see, for example, example 23).

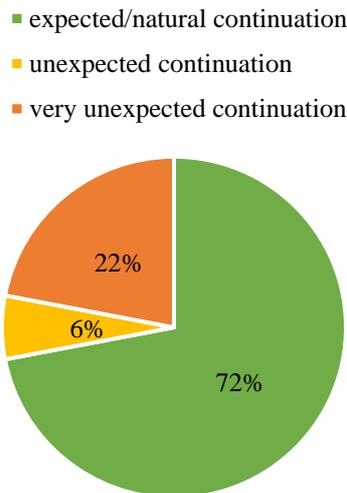


Figure 2: Results of the continuation task (n = 32)

As the results show, ChatGPT had enough Spanish input (training data) to continue most sentences starting with the TI-construction with the most expected following sequence (*pero* ‘but’). ChatGPT’s good performance in this task is simply based on the frequency-based calculation of the most likely continuation. The algorithm seems to recognize this pattern superficially, but it does not always produce adequate continuations, sometimes ignoring the specific context and the immediate QUDs.

In the continuations, ChatGPT shows no variation: all answers start with adversative *pero* ‘but’. However, in our natural corpus data we also find other alternatives that introduce the adversative sentence, such as Sp. *lo que pasa es que* ‘what is happening is that’, *la cuestion es* ‘the question is that’, or *ahora* and *ahora bien* (literally ‘now’ and ‘now well’, also used as an adversative discourse marker). As already observed by other authors (see section 2.2.), LLMs produce less variation and more monotonous texts. In this particular case, ChatGPT always chooses the most frequent adversative conjunction.

The cases in which the continuation is not adversative but affirmative can be explained by the fact that the pattern we expected in this task (TI-construction + adversative sentence) is not the only pattern in which the TI-construction occurs. The construction can also be used for emphatic affirmations that do not entail the rejection of a second QUD or a subset of the QUD (see section 3 and some of the corpus examples in section 4.2). Therefore, in the training data for ChatGPT, most probably not 100% of the cases of the TI-construction were followed by an adversative sentence. However, the specific examples we presented in this task are more natural with an adversative continuation. A possible conclusion from this observation might be that the texts produced by ChatGPT show some inconsistencies in the choice of the appropriate TI-construction pattern according to the specific discourse context.

6 Interpretation task in Spanish (ChatGPT-3.5)

6.1 Methodology

Since the TI-construction followed by an adversative sentence is already routinized in Spanish, the implicit rejection of proposition 2 can also be inferred from the context without having to be explicitly uttered. Therefore, the interpretation task for ChatGPT-3.5 was to interpret the implicit answer to the second proposition, or the part of the QUD that was not explicitly affirmed.

For this task, we used the same set of artificial examples as in the continuation task (section 5) and created 39 dialogues. The prompts included these dialogues and a question about the entailed, second proposition or QUD that was not explicitly answered:

- *¿Cuál es la respuesta implícita a la pregunta “¿Nos casamos?”*? ‘What is the implicit answer to the question “Should we get married?”’
- *¿Cuál sería la respuesta a la pregunta “¿Os habéis separado?” en este diálogo?* ‘What would be the answer to the question “Did you separate?” in this dialogue?’
- *¿Cuál es la opinión de B sobre el sueldo de Manuel?* ‘What is B’s opinion about Manuel’s salary?’

We then classified the answers given by ChatGPT into three categories:

- i. Expected interpretation: ChatGPT gives a negative answer to the question, i.e., it rejects the second proposition or QUD. In this case, the expected pragmatic inference is drawn.
- ii. Unexpected interpretation: ChatGPT gives a positive answer to the question, drawing a mistaken inference that would be blocked by the TI-construction in human conversation.
- iii. No interpretation: ChatGPT states that nothing can be said and that more questions are needed. No inference is drawn at all.

In the previous task, ChatGPT performed quite well in continuing the TI construction with an adversative segment that negates the second proposition (score: 72%). Consequently, one might expect that it is also relatively “easy” to infer adversative continuations that are not explicitly stated (i.e., category i).

6.2 Data

In example (24), ChatGPT draws the intended pragmatic inference and states that the second question is implicitly answered in the negative:

(24a) Spanish (prompt)

¿Cuál es la respuesta implícita a la pregunta “¿Lo has logrado?”

A: *¿Ya has intentado arreglar la nevera? ¿Lo has logrado?*

B: *Intentarlo, lo he intentado.*

‘What is the implicit answer to the question “Did you succeed?”?’

A: Have you tried to fix the refrigerator yet? Did you succeed?

B: As for trying, I have tried.’

(24b) Spanish (ChatGPT’s answer)

La respuesta implícita a la pregunta “¿Lo has logrado?” en este diálogo es negativa o de no éxito. La respuesta de la persona B, “Intentarlo, lo he intentado,” sugiere que B ha intentado arreglar la nevera, pero la construcción gramatical y la falta de una respuesta afirmativa indican que no ha tenido éxito en su intento. En otras palabras, B ha intentado arreglar la nevera, pero no lo ha logrado hasta el momento.

‘The implied response to the question “Did you succeed?” in this dialogue is negative or unsuccessful. Person B’s response, “As for trying, I have tried” suggests that B has tried to fix the refrigerator, but the grammatical construction and the lack of an affirmative response indicate that he has not succeeded in his attempt. In other words, B has tried to fix the fridge, but has not succeeded so far.’

ChatGPT’s answer in (24b) states that the interpretation is based on the grammatical construction (possibly referring to the TI?) and the lack of an affirmative answer. The latter means that even without the TI-construction, this example could probably be interpreted as a negative answer to the second question, simply due to the fact that there is no answer at all regarding the success of the attempt.

Similarly, the expected negative answer to the question in (25) may not be triggered by the TI-construction but by the fact that there is no explicit answer to the question, e.g., out of modesty:

(25a) Spanish (prompt)

¿Se considera B una experta o no?

A: *¡Has leído el libro entero sobre los Reyes Católicos! Supongo que ahora eres una experta en la materia.*

B: *Leerlo, lo he leído, ...*

‘Does B consider herself an expert or not?’

A: You’ve read the whole book on the Catholic Kings! I guess you are now an expert on the subject.

B: As for reading it, I have read it, ...’

(25b) Spanish (ChatGPT’s answer)

La respuesta de la persona B, “Leerlo, lo he leído, ...” sugiere que B ha leído el libro, pero no afirma explícitamente ser una experta en la materia. B más bien da a entender que ha adquirido conocimientos al leer el libro, pero no se autodenomina experta. La respuesta de B es más modesta y no asume un alto grado de conocimiento o experiencia en el tema de los Reyes Católicos.

‘Person B’s response, “As for reading it, I have read it, ...” suggests that B has read the book, but does not explicitly claim to be an expert on the subject. B rather implies that she has gained knowledge from reading the book, but does not call herself an expert. B’s response is more modest and does not assume a high degree of knowledge or expertise on the subject of the Catholic Kings.’

However, in many other examples, ChatGPT incorrectly concludes that the implicit answer is affirmative:

(26a) Spanish (prompt)

¿Cuál sería la respuesta a la pregunta “¿Ahora se puede abrir?” en este diálogo?

A: ¿Ya han arreglado la puerta? ¿Ahora se puede abrir?

B: Arreglarla, la han arreglado...

‘What would be the answer to the question “Can you open it now?” in this dialogue?

A: Has the door been fixed? Can you open it now?

B: As for fixing it, they fixed it...’

(26b) Spanish (ChatGPT’s answer)

La respuesta implícita a la pregunta “¿Ahora se puede abrir?” en este diálogo es afirmativa. La respuesta de la persona B, “Arreglarla, la han arreglado...” indica que la puerta ha sido reparada, lo que sugiere que ahora se puede abrir sin problemas. La respuesta de B confirma que la puerta ha sido arreglada con éxito y que es funcional nuevamente.

‘The implicit answer to the question “Can you open it now?” in this dialogue is yes. Person B’s response, “As for fixing it, they fixed it...” indicates that the door has been repaired, suggesting that it can now be opened without problems. B’s response confirms that the door has been successfully fixed and is functional again.’

As already observed in the continuation task (see example 23 in section 5.2), ChatGPT states that if a door is fixed, you must be able to open it. Real-world exceptions like the one in (26) are not taken into account. A similar case occurs in example (27):

(27a) Spanish (prompt)

Basándonos en este diálogo, ¿sabe B dónde está el libro?

A: ¿Sabes dónde está el libro? ¿Lo has visto?

B: Verlo, lo he visto...

‘Based on this dialogue, does B know where the book is?

A: Do you know where the book is, have you seen it?

B: As for seeing it, I’ve seen it...’

(27b) Spanish (ChatGPT's answer)

Basándonos en el diálogo proporcionado, parece que la persona B sabe dónde está el libro y afirma haberlo visto. La respuesta de B, "Verlo, lo he visto...", indica que no solo ha visto el libro, sino que también está enfatizando esta afirmación, lo que sugiere un alto grado de certeza acerca de la ubicación del libro.

'Based on the dialogue provided, it appears that person B knows where the book is and claims to have seen it. B's response, "As for seeing it, I have seen it..." indicates that not only has she seen the book, but she is also emphasizing this claim, suggesting a high degree of certainty about the location of the book.'

Again, in example (27), ChatGPT strictly follows the assumption that if you have seen something, you surely know where it is. However, human knowledge of the world and everyday experience teach us that there are exceptions to this rule: the answer *verlo, lo he visto* 'as for seeing it, I have seen it' most likely encodes the implicature *pero no recuerdo dónde* 'but I don't remember where'.

Finally, in many cases, ChatGPT does not draw any inference at all but states that more information or another question is needed:⁹

(28a) Spanish (prompt)

¿Cuál es la respuesta implícita a la pregunta "¿Sacas buenas notas?" en este diálogo?

A: ¿Estudias mucho? ¿Sacas buenas notas?

B: Estudiar, estudio...

'What is the implicit answer to the question "Do you get good grades?" in this dialogue?

A: Do you study a lot? Do you get good grades?

B: As for studying, I study'

(28b) Spanish (ChatGPT's answer)

La respuesta implícita a la pregunta "¿Sacas buenas notas?" en este diálogo es incierta. Aunque la persona B reconoce que estudia ("Estudiar, estudio..."), no proporciona una respuesta directa sobre si saca buenas notas o no. La respuesta de B sugiere que está comprometido con el estudio, pero no ofrece información específica sobre sus calificaciones. Por lo tanto, la respuesta implícita a la pregunta "¿Sacas buenas notas?" no está clara en este diálogo.

⁹ As suggested by an anonymous reviewer, we repeated the experiment with this sentence because in the first version, there was no coma between the infinitive and the inflected verb form (i.e., *Estudiar estudio...*). During the revision, we tested all the examples in our sample where the coma was missing, but ChatGPT's answers (on May 12, 2024) did not consistently change when a coma was introduced. The answer stays basically the same, e.g., in example (28).

‘The implicit answer to the question “Do you get good grades?” in this dialogue is uncertain. Although person B acknowledges that he studies (“As for studying, I study...”), he does not provide a direct answer as to whether or not he gets good grades. B’s response suggests that he is committed to studying, but he does not provide specific information about his grades. Therefore, the implicit answer to the question “Do you get good grades?” is unclear in this dialogue.’

In a natural human conversation, the interpretation in (28b) is not impossible either, but unlikely. Probably, most human speakers would try to draw some inference from a deliberately incomplete, evasive answer. This is even more likely if an answer to the question (*¿Sacas buenas notas?*) is avoided on purpose because the negative answer (*No saco buenas notas*) is potentially self-defeating (i.e., a potentially face-threatening speech act).

Interestingly, similar results are obtained when asking ChatGPT in English, using the “*as for + gerund*” construction exactly as provided in the translations in this section. Also in English, an answer like *As for studying, I study...* can be interpreted as evasive, and therefore the same inference can be drawn as in Spanish, although the English construction is probably not that conventionalized. However, when asking ChatGPT in English, the answer remains basically the same (*The implicit answer to the question “Do you get good grades?” in this dialogue is uncertain*). This also holds for the other examples discussed here: in (26) and (27), in English, ChatGPT interprets a false positive answer, but a correct negative answer in (24). This probably indicates that ChatGPT’s (in)ability to draw pragmatic inferences is, to some extent, language independent.

6.3 Results and discussion

ChatGPT mostly failed at the interpretation task. Overall, in most cases, the responses were unexpected (i.e., positive answers), followed by no interpretation at all due to missing information. Only in a very few cases (5%), ChatGPT did give the expected answer, implicitly rejecting the second proposition (see figure 3).

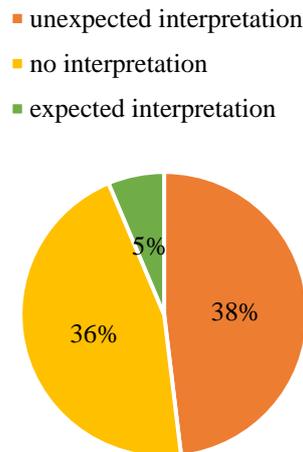


Figure 3: Results of the interpretation task (n = 39)

Thus, while ChatGPT mostly continues the examples we tested with an adversative sentence (see section 5.3), it does not draw the same interpretations when it comes to pragmatic inferences. All examples that ChatGPT failed in the continuation task were failed in the interpretation task, too. However, in addition the algorithm also misinterpreted many examples that, on the contrary, were continued with an adversative sentence in the previous task.

In some examples, the entailment between proposition 1 and proposition 2 is stronger than in other examples, based on a higher frequency of proposition 2 being a consequence of proposition 1. For example, trying to fix something does not always lead to success (example 24), i.e., the verb *intentar* ‘to try’ does not necessarily entail *lograr* ‘to succeed’. It is even likely that the verb *intentar* ‘to try’ is more often followed by linguistic expressions of failure than of success. Similarly, reading a book does not automatically make one an expert in the field (example 25), since it takes much more than a single book, as ChatGPT’s answer also reflects. In both examples, ChatGPT answers as expected and denies the second QUD.

Conversely, if a door has been fixed, the most common result is that it can now be opened (example 26). Also, if you remember seeing a certain book, you usually remember its location (example 27). In this sense, ChatGPT’s answers seem to reflect probabilities: whereas the algorithm denies the second propositions in examples (24) and (25), it strongly affirms them in examples (26) and (27), ignoring abnormal, infrequent situations such as repaired but inoperative doors, or forgetting where you saw a certain book.

7 Conclusion

This exploratory study is concerned with the topicalized infinitive construction in Spanish and Portuguese (e.g., Sp. *comer no come...* ‘as for eating he does not eat’),

which has no literal correlate in English. Our first research question was if LLMs such as DeepL correctly translate this construction from Spanish to Portuguese and *vice versa*, considering that literal translations should be the preferred option. To answer this question, we tested the automated translation of authentic corpus examples.

Our second research question concerned the pragmatics of the TI-construction: we wanted to find out whether ChatGPT recognizes the adversative effect of such constructions. We therefore created a set of Spanish dialogues in which two propositions addressing two entailed QUDs are linked (e.g., A: *Manuel, tu marido, trabaja un montón. Supongo que gana mucho dinero* ‘Manuel, your husband, works a lot. I guess he makes a lot of money.’. B: *Trabajar sí que trabaja mucho...* ‘As for work, indeed he works a lot...’). In our second experiment, ChatGPT was asked to continue these sentences in a natural way. The most expected continuation would be with an adversative sentence (e.g., *pero no gana mucho dinero* ‘but he does not earn much money’).

Finally, in our third experiment, the algorithm was asked about the implicit answer to the second QUD, e.g., what is the indirect answer regarding Manuel’s salary in *Trabajar sí que trabaja mucho...* ‘As for work, indeed he works a lot...’. The expected interpretation would be to infer a contrasting negative answer.

Of course, as a pilot study with a small sample size, our results are preliminary and may be confirmed or refuted by future large-scale research: As for the translation task, we found that DeepL fails at translating the TI-construction from Spanish into Portuguese and *vice versa* since only in 4% to 13% of the cases the TI is translated as such. On the continuation task, ChatGPT performed quite well, continuing 72% of the examples with an expected adversative sentence (*pero...* ‘but’). However, it clearly fails at drawing inferences: in only 5% of the answers, ChatGPT finds that the second proposition is implicitly negated. Table 2 summarizes our findings:

Table 2: Overview of results (score of successful/correct/expected outcome)

✗	DeepL	Translation task	Sp. < > Pt.	4%–13%
✓	ChatGPT	Continuation task	Sp.	72%
✗	ChatGPT	Interpretation task: pragmatic inferences	Sp.	5%

The three different tasks show interesting results that can be explained by the mechanisms underlying LLMs:

DeepL’s bad performance in the translation task is surprising at first sight, since the literal translation with an infinitive would not be only the “simplest” solution (at least from a human perspective) but also the most adequate translation, given that the TI-construction exists and is used equivalently in Spanish and Portuguese. At a second glance, this result needs to be explained by English contamination: as shown in our sample, DeepL surely translates via English as the pivot language. In this respect, our pilot study confirms the observation already made by Šorak (2020). English contamination is problematic insofar as Spanish and

Portuguese are closely related languages, and translation via English, a language that lacks the TI-construction, results in a remarkable loss of stylistic idiosyncrasies.

This is even more problematic if we consider that information structure patterns in general seem to be altered by DeepL: Non-canonical word order in Spanish and Portuguese carries relevant information about how to interpret a so marked constituent of a sentence (topic or focus). Changing this word order in automated translations results in more monotonous sentences that lack syntactic topic and focus marking. For example, also the sentence-final position of the focused subject in Spanish and Portuguese sentences as well as clitic-left-dislocations tend to get lost in translation from Spanish to Portuguese and *vice versa*. Of course, further research on a broader sample than the one tested here is needed. A quick spot check in ChatGPT suggests that it may be slightly better in this regard. In order to improve automated translation in the future, we believe it would be crucial to eliminate English contamination, especially in language pairs such as Spanish-Portuguese.

On the contrary, ChatGPT's good score on the continuation task confirms that the mechanism behind this LLM is successful: it is a pure probabilistic calculation of the most likely following segments (*pero* 'but'). Nevertheless, not all continuations were adequate in the sense that not all answered the immediate QUD in the given context, but just gave a random sentence introduced by *pero* 'but'.

The results of the third task – the interpretation task – are somewhat inconsistent with the continuation task: while ChatGPT continues most TI-constructions with an adversative sentence, it fails to infer these adversative implicatures in exactly the same examples. In other words: ChatGPT is good at mimicking common speech patterns, but has no idea what it is talking about. This has already been pointed out by many authors: e.g., Bender, Gebru, McMillan-Major and Shmittchel (2021: 617) state that LLM are “stochastic parrots”. Furthermore, as such, ChatGPT seems to perform worse in examples that describe less frequent and unlikely situations, such as having fixed something but still not being able to use it. On the contrary, the algorithm performs better at interpreting and predicting probably more frequent (and thus more probabilistic) adversative associations between propositions, such as *intentar, pero no lograr* ‘to try, but not to succeed’.

In future research, our findings could be refined by a new study that tests TI-constructions after explicit instruction on how to interpret the structure. Most likely, ChatGPT's performance would improve after few-shot prompting, i.e., after giving specific instructions and examples, as demonstrated in other studies (Kim, Taylor and Kang 2023; Ruis, Khan, Biderman et al. 2022). After all, humans also need to learn how to draw pragmatic inferences. Nevertheless, we think that, unlike humans, LLMs will not use indirect language by themselves, i.e., it is quite unlikely that ChatGPT itself would produce a TI-construction without continuation (as in the third task) to avoid an explicit answer that could be face-threatening (e.g., in our examples: when admitting bad grades, a low salary, or rejecting a marriage proposal). On the contrary, the answers provided by the algorithm are overabundant

and tend to rather repeat information unnecessarily. For now, it seems that ChatGPT is not “aware” of the fact that – in certain contexts – giving a shorter, indirect answer to implicitly convey more meaning is even more efficient and more polite.

Acknowledgments

I would like to thank the organizers and participants of the *1st International Conference on “Automated texts In the ROMance languages” (Ai-ROM)* (28–29 September 2023) as well as the two anonymous reviewers for their comments, objections, and feedback, which consistently improved this paper. Of course, any remaining errors and misinterpretations are mine.

References

- Barattieri di San Pietro, Chiara & Frau, Federico & Mangiaterra, Veronica & Bambini, Valentina. 2023. The pragmatic profile of ChatGPT: Assessing the communicative skills of a conversational agent. *Sistemi Intelligenti(2)*. 379–400.
- Bastos, Ana Cláudia P. 2001. *Fazer, eu faço! Topicalização de constituintes verbais em português brasileiro*. Campinas, São Paulo: Universidade Estadual de Campinas (Master’s Thesis).
- Bechara, Evanildo. 2009. *Moderna Gramática Portuguesa*. 37th edn. Rio de Janeiro: Nova Fronteira.
- Bender, Emily M. & Gebru, Timnit & McMillan-Major, Angelina & Shmitchell, Shmargaret. 2021. On the dangers of stochastic parrots: Can Language Models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. New York: ACM. <https://doi.org/10.1145/3442188.3445922> (last accessed 02.02.2024)
- CDH = Instituto de Investigación Rafael Lapesa de la Real Academia Española. 2013. *Corpus del Nuevo diccionario histórico del español*. <http://web.frl.es/CNDHE>.
- CdP = Davies, Mark & Ferreira, Michael. 2006. *Corpus do Português: 45 million words, 1300s-1900s*. <http://www.corpusdoportugues.org/hist-gen>
- COSER = Fernández-Ordóñez, Inés. 2005–. *Corpus Oral y Sonoro del Español Rural*. www.corpusrural.es
- De Cesare, Anna-Maria. 2021. Répétitions et variations des textes générés: Une analyse linguistique basée sur un corpus d’articles financiers rédigés en français. *CHIMERA: Romance Corpora and Linguistic Studies* 8. 79–108. <https://doi.org/10.15366/chimera2021.8.004> (last accessed: 29.01.2024)
- De Cesare, Anna-Maria. 2023. Assessing the quality of ChatGPT’s generated output in light of human-written texts: A corpus study based on textual parameters. *CHIMERA: Romance Corpora and Linguistic Studies* 10. 179–210. <https://revistas.uam.es/chimera/article/view/17979> (last accessed: 29.01.2024)

- Enrique-Arias, Andrés & Gerhalter, Katharina. Submitted. *Morir morirás* ‘for certain you will die’. Strategies for translating the Hebrew infinitive absolute in medieval and early modern Spanish biblical translations.
- Escandell Vidal, María V. 1991. Sobre las reduplicaciones léxicas. *LEA: Lingüística española actual* 13(1). 71–86.
- Gerhalter, Katharina. In press. *Escrever não escrevo, mas ler um livro, ou um jornal, uns versos, leio*. A corpus-linguistic approach to topicalized infinitives in Portuguese. In Calderón Campos, Miguel & Vaamonde, Gael (eds), *Linguistic Corpora and Big Data in Spanish and Portuguese*. Berlin: De Gruyter.
- Gerhalter, Katharina. In preparation. A diachronic corpus study on topicalized infinitives in Romance languages.
- Hadfield, Jeremy. 2022. *Why Large Language Models will not understand human language*. <https://jeremyhadfield.com/why-llms-will-not-understand-language> (last accessed: 29.01.2024)
- Hein, Johannes. 2020. *Verb Doubling and Dummy Verb. Gap Avoidance Strategies in Verbal Fronting*. Berlin: De Gruyter.
- Hein, Johannes. 2021. Verb movement and the lack of verb-doubling VP-topicalization in Germanic. *The Journal of Comparative Germanic Linguistics* 24(1). 89–144.
- Hu, Jennifer & Floyd, Sammy & Jouravlev, Olessia & Fedorenko, Evelina & Gibson, Edward. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4194–4213. <https://aclanthology.org/2023.acl-long.230> (last accessed: 29.01.2024)
- Kim, Zae M. & Taylor, David E. & Kang, Dongyeop. 2023. “Is the Pope Catholic?” Applying chain-of-thought reasoning to understanding conversational implicatures. *arXiv*. <https://arxiv.org/pdf/2305.13826> (last accessed: 29.01.2024)
- Kocoń, Jan & Cichecki, Igor & Kaszyca, Oliwier & Kochanek, Mateusz & Szydło, Dominika & Baran, Joanna & Bielaniec, Julita & Gruza, Marcin & Janz, Arkadiusz & Kanclerz, Kamil et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* 99. 101861. <https://doi.org/10.1016/j.inffus.2023.101861> (last accessed: 29.01.2024)
- Leonetti, Manuel & Escandell Vidal, María V. 2021. La estructura informativa. Preguntas frecuentes. In Leonetti, Manuel & Escandell Vidal, María V. (eds), *La estructura informativa*, 15–181. Madrid: Visor Libros.
- Meier-Vieracker, Simon. 2024. Automated football match reports as models of textuality. *Text & Talk*. <https://doi.org/10.1515/text-2022-0173> (last accessed: 29.01.2024)
- Muñoz Pérez, Carlos & Verdecchia, Matías. 2022. Predicate doubling in Spanish. On how discourse may mimic syntactic movement. *Natural Language & Linguistic Theory* 40. 1159–1200.

- Narbona Jiménez, Antonio. 2015. *Sintaxis del español coloquial*. Sevilla: Editorial Universidad de Sevilla.
- Nieto García, Paola & Cases Berbel, Elke. 2022. Traducción de DeepL de los sujetos nulos de un texto literario hacia lenguas románicas pro drop y no pro drop. *CLINA Revista Interdisciplinaria de Traducción Interpretación y Comunicación Intercultural* 7(2). 41–59.
<https://doi.org/10.14201/clina2022724159> (last accessed: 29.01.2024)
- Piantadosi, Steven. 2024. Modern language models refute Chomsky's approach to language. In Gibson, Edward & Poliak, Moshe (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett*. Berlin: Language Science Press.
<https://langsci-press.org/catalog/book/434> (last accessed: 29.01.2024)
- Qiu, Zhuang & Duan, Xufeng & Cai, Zhenguang G. under review. Pragmatic implicature processing in ChatGPT. *Cognition*. Preprint:
<https://doi.org/10.31234/osf.io/qtbh9> (last accessed: 29.01.2024)
- Reich, Uli. 2011. Frontalizaciones de la semántica verbal en español y portugués. (Paper presented at the *18th Deutscher Hispanistentag*, Passau, 23–26 March 2011).
- Ruis, Laura & Khan, Akbir & Biderman, Stella & Hooker, Sara & Rocktäschel, Tim & Grefenstette, Edward. 2022. Large language models are not zero-shot communicators. *arXiv*. <https://arxiv.org/pdf/2210.14986.pdf> (last accessed: 29.01.2024)
- Šorak, Vanessa. 2020. *Die Kontaminierung maschineller Übersetzungsprozesse durch das Englische*. Heidelberg: Ruprecht-Karls-Universität Heidelberg (Master's Thesis).
- Srivastava, Aarohi & Rastogi, Abhinav & Rao, Abhishek & Shoeb, Abu Awal Md & Abid, Abubakar & Fisch, Adam & Brown, Adam R. & Santoro, Adam & Gupta, Aditya & Garriga-Alonso, Adrià et al. 2023. Beyond the imitation game. Quantifying and extrapolating the capabilities of language models [version: June 2023]. *Transactions on Machine Learning Research* 5.
<https://jmlr.org/tmlr/papers> (last accessed 29.01.2024)
- Stark, Elisabeth. 1997. *Vorstellungsstrukturen und „topic“-Markierung im Französischen. Mit einem Ausblick auf das Italienische*. Tübingen: Narr.
- Valenzuela, Javier & Hilferty, Joseph & Garachana-Camarero, Mar. 2005. On the reality of constructions. The Spanish reduplicative-topic construction. *Annual Review of Cognitive Linguistics* 3. 201–215.
- Vicente, Luis. 2007. *The Syntax of Heads and Phrases. A Study of Verb (Phrase) Fronting*. Leiden: Leiden University (Ph.D. Dissertation)
- Vigier-Moreno, Francisco J. & Pérez-Macías, Lorena. 2022. Assessing neural machine translation of court documents. A case study on the translation of a Spanish remand order into English. *Revista de Llengua i Dret* 78. 73–91.
<https://doi.org/10.2436/rld.i78.2022.3691> (last accessed 29.01.2024)
- Zheng, Zilong & Qiu, Shuwen & Fan, Lifeng & Zhu, Yixin & Zhu, Song-Chun. 2021. GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational Reasoning. In Zong, Chengqing & Xia, Fei & Li, Wenjie &

Navigli, Roberto (eds), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2074–2085. Stroudsburg, PA: Association for Computational Linguistics. <https://aclanthology.org/2021.findings-acl> (last accessed 29.01.2024)