Crowdsourcing Veridicality Annotations in Spanish: Can Speakers Actually Agree?

Teresa Martín Soeder
Saarland University / Saarbrücken, Germany
temart@lst.uni-saarland.de

Abstract

In veridicality studies, an area of research of Natural Language Inference (NLI), the factuality of different contexts is evaluated. This task, known to be a difficult one since often it is not clear what the interpretation should be Uma et al. (2021), is key for building any Natural Language Understanding (NLU) system that aims at making the right inferences. Here the results of a study that analyzes the veridicality of mood alternation and specificity in Spanish, and whose labels are based on those of Saurí and Pustejovsky (2009) are presented. It has an inter-annotator agreement of $AC_2 = 0.114$, considerably lower than that of de Marneffe et al. (2012) ($\kappa = 0.53$), a main reference to this work; and a couple of mood-related significant effects. Due to this strong lack of agreement, an analysis of what factors cause disagreement is presented together with a discussion based on the work of de Marneffe et al. (2012) and Pavlick and Kwiatkowski (2019) about the quality of the annotations gathered and whether other types of analysis like entropy distribution could better represent this corpus. The annotations collected are available at https://github.com/narhim/veridicality_spanish.

1 Introduction

Often when hearing an utterance we try to assess whether the information conveyed is likely to be truthful or not, that is, if it corresponds to actual situations in the real world (Saurí and Pustejovsky, 2009). Furthermore, as speakers or authors, we normally seek to convey what we know about the **truthfulness** or **factuality** of the events conveyed. For simplicity, here events are just considered as *anything that happens or is* like ”being tall” or ”having read a book”.

In the realm of linguistics, several features lead us to make the correct inferences about the factuality of an event, and comprehending them is key for building any system that aims at understanding human language. For example, the presence of the negation adverb ”not” in ”Pedro has not done the laundry”, leads us to infer that the event ”Pedro has done the laundry” did not happen, unless we know something about the speaker or Pedro that makes us think otherwise. Furthermore, in ”Pedro could have done the laundry” the modal verb ”can” makes the event a possibility. This kind of analysis is what is called a veridicality study, more specifically, **veridicality** is an area of research within natural language inference (NLI) and theoretical linguistics that studies the truth value of a proposition or event in a specific context (Giannakidou, 2014; Giannakidou and Mari, 2015).

As to NLI, it is a branch of natural language understanding (NLU) with its main task being entailment classification, that is, as it has been done above, to classify the relationship between two sentences, a premise, and a hypothesis, by picking a label from a usually small set of labels like {*entailment, neutral, contradiction*} (Williams et al., 2017) or {*yes, unknown, not*}, depending on how the task is defined. So for example, we can classify the relationship between the premise ”Pedro has not done the laundry” and the hypothesis ”Pedro has done the laundry” as a contradiction or not.

Here a study of veridicality judgments in Spanish is presented. Specifically, the goal is to analyze how mood alternation, in other words, the possibility of using a verb either in indicative or subjunctive mood; and the specificity of the syntactic subject, that is, the identifiability of the referent in the discourse universe (Caudet, 1999), affect factuality judgments about an event. This goal is realized in the following research questions:

Proceedings of the Student Research Workshop associated with RANLP-2023, pages 68–77, Varna, Bulgaria, Sep 4–6, 2023.
https://doi.org/10.26615/issn.2603-2821.2023_008

RQ1.- In a complex sentence, how does the mood alternation of the embedded verb that occurs due to the negation of the main or matrix verb affect the factuality value of the embedded event?

RQ2.- How does an individual subject affect the factuality judgment of the event?

RQ3.- How does a subject that refers to a collective entity like an institution, affect the factuality judgment of the event?

To answer these questions a crowdsourcing experiment was run on Toloka (Pavlichenko et al., 2021) in which annotations from linguistically naive native Spanish speakers were gathered for a corpus specifically designed. The corpus and the annotations are publicly available and their analysis is shown here.

Next, Section 2 introduces the main concepts used here and presents the most important references to this work. Then, Section 3 explains how the corpus was designed and how the annotations were gathered. After that, Section 4 presents the statistical and linguistic analysis of the annotations gathered, Section 5 discusses the main issues seen throughout the study, and, finally, Section 6 answers the research questions and proposes some lines of future work.

2 Background

2.1 Veridicality and Factuality

Let us consider examples (1a) to (1d), where we have events that are intrinsically related. If we were to do an NLI study with these examples, we could directly study the *thruthfulness* of each of them as a single event, i.e., we could study the **factual nature** of each example towards the real world or the events in the discourse (Saurí and Pustejovsky, 2009). Another option would be to study the factual nature of the event *Anna's father has arrived* in the different contexts in which is presented: standing completely on its own (1a), or as part of a complex event (1b) to (1d). In this case, the goal would be not to understand the factuality of *Anna's father has arrived*, but rather to understand how its factuality changes when the event is embedded under an epistemic verb (1b), a verb of believe (1c), and a verb of speech (1d). The former case is a factuality study and examples of it are the XNLI corpus (Conneau et al., 2018) and the work of Pavlick and Kwiatkowski (2019). The latter is a **veridicality study**, as the study of Ross and Pavlick (2019) and the experiment presented here.

(1) a. Anna's father has arrived.

b. John knows that Anna's father has arrived.

c. John believes that Anna's father has arrived.

d. John says that Anna's father has arrived.

2.2 Lexical and Pragmatic Approach

When designing an NLI study there are two main possible approaches: lexical and pragmatic. In the first case, the aim is to model the aspects of a sentence semantics (Ross and Pavlick, 2019), and thus, its representation can be derived from the lexicon and is independent of context, which mean the omission of world knowledge. To follow this approach, annotations must be gathered from linguistic experts. Examples of corpora with this approach are the FactBank corpus Saurí and Pustejovsky (2009) in English, and the SenSem (Fernández-Montraveta and Vázquez, 2014) and TAGFACT (Fernández Montraveta et al., 2020) corpora in Spanish.

As to the pragmatic approach, which is used here, it aims at modeling a representation of the sentence that considers the communication intent for that sentence in a specific context, that is, a goal-directed representation of a sentence within the context it was created (Ross and Pavlick, 2019). To obtain such a representation one needs to consider world knowledge and embrace uncertainty (de Marneffe et al., 2012). Furthermore, to follow this approach, annotations must be gathered from linguistically naive workers. Examples of studies that follow this approach are de Marneffe et al. (2012); Conneau et al. (2018); Ross and Pavlick (2019) and Pavlick and Kwiatkowski (2019).

2.3 Mood Alternation in Spanish

In its most basic definition mood is said to be the grammaticalization of modality (Lyons, 1995; Sánchez-Jiménez, 2011), and thus it has been traditionally related to the speaker's attitude towards an utterance (Lyons, 1995; Real Academia Española, 2011). Furthermore, since the commitment of the speaker usually takes form in different degrees (Lyons, 1995), in most languages, mood takes form in different subcategories. For Spanish, nowadays most of grammarians agree on the existence of three subcategories of mood: indicative, subjunctive, and imperative. Only indicative and subjunctive are relevant for our purposes here.

One of the ways in which the different mood categories are distinguished is based on their syntactic behavior. Specifically, authors often talk about a dependent and an independent mood (Real Academia Española, 2011), the first one being the one that requires a grammatical inductor to appear, and the second being the one that does not need any grammatical elements to appear in the sentences. This distinction mostly correlates with the subjunctive and the indicative moods, that is, normally, for a verb to be in the subjunctive mood there must be a grammatical element that induces it.

Usually, the induced mood is *mandatory*, that is, using the verb in a different mood category is not accepted. But there are cases in which a different mood category, in most cases the indicative, is accepted and this is what is called **mood alternation**, one of the veridicality contexts analyzed here.

Specifically, the focus here lays on the mood alternation that occurs in the embedded predicate of a complex sentence due to the negation of the main or matrix verb, as in example (2), where due to the presence of the negation adverb *no* (not), the embedded verb *tener* is allowed to appear both in the subjunctive (example (2b)) and in the indicative (example (2a)). Since there is no direct way of translating the mood differences into English, here, as in Faulkner (2021), the translations are identical.

In this case, the difference in the interpretation between indicative and subjunctive is interpreted in terms of old and new information. That is, when the speaker chooses to use the subjunctive mood it is understood that the embedded event is already part of the common ground. Contrary to this, when using the indicative, the embedded event is presented as new information (Mejías-Bikandi, 1998; Real Academia Española, 2011; Faulkner, 2021). Consequently, the event *el país tenía problemas económicos* (the country had economic problems) is presented as part of the common ground in (2b), and as new knowledge in (2a). Because it was assumed that speakers associate different factuality values with old and new information, it was expected that mood alternation would alter the factuality of the embedded event.

(2) a. El presidente no
the.M.SG president.M.SG not
dijo que el
say.PST.PFV.IND.3SG that the.M.SG
país **tenía**
country.M.SG **have.PST.IPFV.IND.3SG**
problemas económicos.
problem.M.PL economic.M.PL
"The president didn't say that the country **had** economic problems."

b. El presidente no
the.M.SG president.M.SG not
dijo que el
say.PST.PFV.IND.3SG that the.M.SG
país **tuviera**
country.M.SG **have.PST.IPFV.SBJV.3SG**
problemas económicos.
problem.M.PL economic.M.PL
"The president didn't say that the country had economic problems."

2.4 Specificity

Following Caudet (1999), here specificity is considered as the identifiability of the referent in the discourse universe and is shown, for example, in the amount and type of information used in the referral expression. So when referring to Olaf Scholz, the current German chancellor, we could use the expression "the German chancellor" or "the chancellor". Assuming the reference is successful in both cases, in the first case the speaker uses more information because she assumes that in the mind speaker, there is more than one chancellor with equal prominence, and thus more information is needed to ensure the right one is chosen. In the second case, no additional information is needed because the speaker assumes only Olaf Scholz is prominent in the mind of the speaker.

Here, the specificity of the subject is manipulated by changing the type of information by having individual vs. collective nouns as subjects. With this, we are manipulating the number of individual entities the subject refers to in singular. In the first case, with an individual noun like *el presidente* (the president) we are referring to one single entity, whereas in the second case with a collective noun like *el gobierno* (the government) we are referring to a set of entities. Because I assumed that there is a different factuality associated with individual and collective nouns, it was expected that there could be a veridicality effect, but not a strong one.

2.5 Previous Work

An important reference is that of de Marneffe et al. (2012), which aimed at identifying some of the linguistics and contextual factors that shape readers' veridicality judgments. To fulfill this goal they crowdsourced annotations on a part of the FactBank corpus and built a system for veridicality assessment. For our purposes, the most important part of their work is the consideration of the possible occurrence of **label split**, that is, that for some premise-hypothesis pairs, there is not just one ground truth and therefore label, associated with them, but at least two, which they concluded from the analysis of the **agreement patterns**, that is, of how the votes for each label are distributed in each pair.

Another relevant work is Pavlick and Kwiatkowski (2019), whose goal was to determine whether the disagreement often seen in NLI datasets is noise or an important reproducible signal. To do so they gathered factuality judgments on 500 pairs, with 50 annotators per pair, of these, 496 pairs with a mean of 39 workers were left to analyze. The results showed that for 20% of the pairs a second ground truth or label can be associated with them, which they blame on **inherent disagreement**.

In Spanish, the main related works are the following corpora: XNLI Conneau et al. (2018), SenSem (Fernández-Montraveta and Vázquez, 2014) and TAGFACT (Fernández Montraveta et al., 2020). XNLI is a multilingual corpus that follows the premise-hypothesis design, but the other two do not. Thus the corpus presented here covers the lack of Spanish corpora in the form of premise-hypothesis pairs and, as far as I know, is the only dataset that focuses on specific phenomena. This, together with the fact that as far as I know the inter-annotator agreement score used here has not been used in any previous NLI corpora, forces us to take any comparisons with previous work skeptically.

3 Corpus and Annotation Process

The first step for creating the corpus was defining the experimental design. To do so, each of the research questions was set as one experimental condition: negation, individual, and collective. Then the negation condition was divided into three categories: baseline, indicative, and subjunctive. The first refers to the case where there is no mood inductor, as in *El presidente dijo que el país tenía problemas económicos* (The president said that the country had economic problems). For both the indicative and the subjunctive categories we have the mood inductor *no* (not), but on the former the embedded verb is in the indicative mood, as in (2a), and on the latter, the verb is in the subjunctive mood, as in (2b). Finally, these three negation categories were crossed with the specificity conditions, that is, with the individual and collective conditions.

Once the design was defined, the pairs were created. 30% of them were written manually and the rest were based on different corpora. Specifically, possible premises in the indicative category were extracted with the help of Linguakit (Gamallo et al., 2018) from the following corpora: a section of the Davie's Corpus del Español (Davies, 2016), the Old News Corpus for Spanish (Kaggle, 2018), *El Quijote* by Miguel de Cervantes (as found in Dario (2017)), the XNLI corpus, and the United Nations corpus in Spanish for the years 2000, 2001, 2002, and 2003 (Eisele and Chen, 2010). After that, some small modifications like reference resolution and reducing the number of words were done, hypotheses were extracted and pairs were modified according to the experimental design. Finally, each pair was automatically annotated with additional information that could be used to later model the results.

To annotate the corpus the labels displayed in Figure 1 were used. These labels correspond to the set from Saurí and Pustejovsky (2009) minus *certain but unknown output* (CTu), as in de Marneffe et al. (2012), and although they mapped them to the traditional square of opposition, here the labels are presented in an ordered linear scale of factuality. Furthermore, given that the acceptability of mood alternation is not always certain, the label ”not a sentence” (NaS) was added. Since does not fit within the scale, it is presented outside of it. Consequently, the final set of labels is: certainly yes (CT+), probably yes (PR+), possibly yes (PS+), unknown or uncommitted (Uu), possibly not (PS-), probably not (PR-), certainly not (CT-), not a sentence (NaS).

The experiment was run on the platform Toloka (Pavlichenko et al., 2021). To select the workers two main criteria were used: language and country. They were required to have set up Spanish as a language and their IP address from a country where Spanish is an official language or an important minority one. In the beginning, all annotators under these criteria were eligible, but then this was reduced to the top 30% of annotators. They were paid $\$0.433$ per set of pairs, which consisted of no more than 10 pairs. Once all the annotations were gathered, pairs for which one worker or more used the label NaS were removed, and if an annotator had labeled more than 1 pair within a single combination of experimental conditions, all his annotations in that combination were removed. This left a total of 477 pairs and 7 annotators per pair.

The task was designed as in de Marneffe et al. (2012): Given a context (the premise), workers had to label the factuality of the event (hypothesis) by choosing one of the labels in Figure 1 from a drop list.

4 Analysis of Annotations

Overall Distribution. As we can see in Figure 2, the distribution of label counts is negatively skewed, that is, there is a clear preference for the positive labels, even if more than half of the corpus sentences are negated. Furthermore, the frequencies for probability and possibility, with their respective + and - signs, are almost identical. This points to a likely confusion for the annotators between probability and possibility. Lastly, we have that for 42.348% of the pairs annotators could not agree upon one label, which suggests a considerable lack of agreement.

Inter-Annotator Agreement Score for Different Subsets	
Subset	AC_2
ALL	0.114
Baseline	0.194
Indicative	0.070
Subjunctive	0.085
saber (to know)	0.170
olvidar (to forget)	0.181
creer (to believe)	0.131

Table 1: Inter-annotator-agreements scores for the whole corpus and different subsets.

Inter-Annotator Agreement Scores. Given that the labels used are not nominal, but ordinal; and the highly skewed distribution seen in Figure 2, here I follow Vanacore and Pellegrino (2022) and computed the inter-annotator agreement score as measured by Gwet's AC_2 (Gwet, 2014). This yielded a value of 0.114, which is barely within the range of slight agreement (Shrout, 1998). Given this, I decided to explore the value of this score in different subsets of the data and the most informative values are in Table 1. There we see that there is a considerable difference between the baseline and two mood alternation categories, but barely between the latter. In addition, we have the scores for the subset of pairs where *saber* (to know) is the matrix verb, and where we have *olvidar* (to forget) as the matrix verb. The agreement in the subsets is quite close, despite being very different in its size (120 pairs for the first one, 24 for the latter), which suggests that agreement depends not on the frequency of the matrix, but on the matrix itself. Further proof of this is the fact that agreement for the subset of *creer* (to believe) is quite lower than the other two, despite having double the pairs than the *olvidar* (to forget) subset.

Model Fitting. A cumulative link mixed model (CLMM) with a logit link was fitted to the whole dataset by using the R software, specifically the ordinal package (Christensen, 2018). This

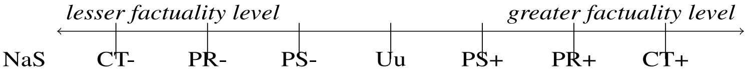

Figure 1: Ordered representation of the labels used for annotating the corpus. Each label stands for: certainly yes (CT+), probably yes (PR+), possibly yes (PS+), unknown or uncommitted (Uu), possibly not (PS-), probably not (PR-), certainly not (CT-), not a sentence (NaS). The latter doesn't fit in the scale, thus it's presented outside of it.

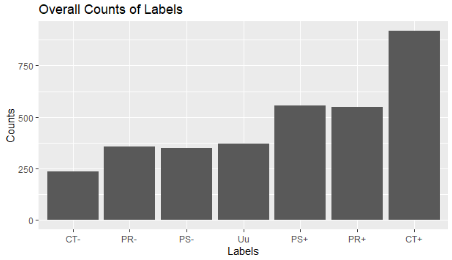

Figure 2: Overall distribution of the proportion of labels used by annotators.

type of model was chosen to reflect the ordered nature of the labels. Mood categories were set as predictors, labels as outcome, and annotator and pair as random variables. Results showed that both the indicative and subjunctive categories are significantly different from the baseline, which is consistent with the agreement scores. The coefficients for these two categories are both negative, rather small in value (< 0.5), and barely different from each other (< 0.1). To assess whether this difference is significant or not, a CLMM was fitted to just these two categories and the results show that in general there is no significant difference between the verb being in the indicative or in the subjunctive mood. Furthermore, the possibility of adding two other predictors, specificity conditions and matrix, to the overall model was considered. In the first case, it was proved that the specificity conditions are not informative. On the second one, the value of the conditional Hessian increased so much (from 1.4×10^2 to 3.3×10^5), that the model was disregarded. Since it was suspected that this increase could be due to the uneven distribution of matrices, the same two models were fitted to the subset corresponding to the 5 most frequent matrices (frequencies ranging from 36 to 102) and it was observed that the conditional Hessian values are much closer (1.4×10^2 to 2.6×10^3) and there are more significant effects for the model with matrix as a predictor, than for the exact same model for the whole dataset. In addition, for the model fitted with both predictors fitted to this subset, the difference between the coefficients for the indicative and subjunctive increased to 0.61. Lastly, when fitting the model with both predictors to the indicative and subjunctive pairs of this small subset, a small ($p = 0.0347$) significant effect for the subjunctive category was found, although not for the specific matrices.

Agreement Patterns. As in de Marneffe et al. (2012), the distribution of the votes for each label in each pair, which can be seen in Figure 3, was analyzed. Although several patterns occur less than 25 times, there are a few that have a non-neglectable frequency. Particularly, $[3, 2, 1, 1]$ and $[2, 2, 1, 1, 1]$ have a frequency of 112 and 134 (23.480% and 28.092%) respectively, which suggests that there are pairs for which disagreement is not an error but rather their underlying truth, even if they cannot be matched to an exact label split. In other words, Figure 3 shows that there is inherent disagreement for $\sim 50\%$ of the corpus.

Manual Analysis. An exploratory manual analysis of the annotations showed that there are other veridicality contexts and other factors that can at least partially explain the lack of agreement found. The two most salient factors are world knowledge and the presence of modal verbs in either the main or the embedded predicate. In support of the former, we have example (3), which was annotated as CT+ by 4 workers, even its variants in the indicative and the subjunctive conditions had the same label with 6 votes for each of them. This is because the factuality of the hypothesis, shown in (3b), cannot be easily negated, even in the presence of more than one veridicality context, since its often considered a *universal truth*. In support of the latter, we have the fact that for 15 out of 30 pairs that have the modal verb *deber + infinitive* (to must + infinitive)[1] in the embedded predicate, there is no agreement

[1] In Spanish there are two constructions with *deber*: deber + infinitive and deber + de + infinitive.

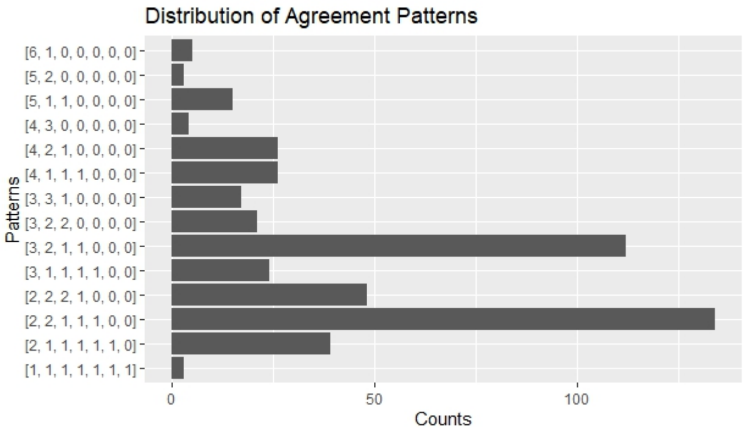

Figure 3: Distribution of agreement patterns, that is of how the annotators' votes are distributed for each pair. Therefore each number in the pattern represents the number of votes for a label (which one depends on the pair) and the counts stand for the number of pairs that have such agreement pattern.

upon one label, proportion $\sim 8\%$ higher than for the whole corpus.

(3) a. Algunos delegados
some.M.PL representative.M.PL
gubernamentales saben
governmental.M.PL know.PRS.IND.3SG
que también las mujeres
that also the.F.PL woman.F.PL
son seres humanos.
be.PRS.IND.3PL being.M.PL human.M.PL

Some governmental representatives know that women are also human beings.

b. Las mujeres son
the.F.PL woman.F.PL be.PRS.IND.3PL
seres humanos.
being.M.PL human.M.PL
Women are human beings.

5 Discussion

The results of this study have shown that there is a tendency for annotators to use positive labels, even when most of the pairs of the corpora are negated. Furthermore, the inter-annotator agreement score, $AC_2 = 0.114$, is rather low since it is barely within the range of slight agreement. Therefore, it is important to discuss what could have caused such a lack of agreement.

Uma et al. (2021) defines five sources of disagreement: errors and interface problems, annotation scheme, ambiguity, item difficulty, and subjectivity. Although the first factor cannot be completely disregarded, given the persistence of some quite divided agreement patterns seen in Figure 3 and that the interface was a simple drop list, the higher level of disagreement cannot entirely be blamed on this first factor.

As to the annotation scheme, some improvements could most certainly be implemented. Specifically, implementing training and testing. This was not done in fear of leading annotators towards concrete labels, but after encountering the work of Nie et al. (2020), where they used carefully crafted training and testing that did not fully prevent disagreement, I understood it could be possible, even if not easy. Also, given that the specific criteria for quality control, like what exactly annotations done too fast look like, more carefully defined criteria would clear out results. In addition, since the model fittings for subsets with more even distributions concerning the matrices showed better results, a more balanced corpus could yield more informative CLMMs, but there is no reason to believe that it would improve agreement. Lastly, given the negatively skewed distribution for the label counts, the assumptions made about the veridicality of negation need to be revised.

Regarding the ambiguity of the relation between the different premises and their hypothesis, the two factors mentioned in the manual analysis (world knowledge and modal verbs) and the highly frequent divided agreement patterns shown in Figure 3, suggest that there is not always one clear label for a pair. This supports the existence of *inherent* disagreement between annotators, although more data is needed to confirm it. Furthermore, the fact that these annotations are done from a pragmatic perspective and that mood alternation is a pragmatic phenomenon, also increases the uncertainty, and therefore ambiguity, of the annotations.

Concerning the fourth possible cause of disagreement, item difficulty is here a certain cause of disagreement. Firstly, in the manual analysis, it was demonstrated that there are different factors to be considered when given a factuality judgment, mainly world knowledge. Secondly, previous work has shown NLI annotations to be difficult (Pavlick and Kwiatkowski, 2019; Uma et al., 2021).

As to the last factor for disagreement, subjectivity, it also influences the results presented here. Although, to my knowledge, there is no previous work that supports this as cause for disagreement in NLI annotations, the analysis of example 3 shows that world knowledge influences speakers' judgments, consequently making annotations dependent upon annotators' knowledge and point of view.

Now that it has been explained what caused disagreement in these studies, the question is if an inter-annotator agreement score, let it be AC_2 or any other, can reflect the nature and quality of the annotations gathered. The simple answer is no, at least not entirely. As stated in Gwet (2014), inter-annotator agreement scores reflect how much the annotations change when small adjustments in the annotators like replacing a number of them are made, that is, it is a measure of data reproducibility based on the individual annotators. However given that it has been proven that these annotations are highly dependent on the speaker, measuring the reproducibility of the data based on such small variations is misguided and different evaluation scores are needed.

6 Conclusions

Based on the results presented here we can conclude that the specificity of the subject, defined in terms of the identifiability of the referent in the discourse universe and manipulated by having individual and collective nouns does not have a significant effect on the factuality of the embedded predicate in complex sentences, or in other words, individual vs. collective subjects are non-veridical contexts concerning the embedded predicate.

As to the effect of mood alternation due to the negation of the matrix verb, that is, when due to negative adverb *no* (not) modifying the main verb of a complex sentence the subjunctive mood is induced in the embedded predicate but the indicative is also accepted; there is overall a significant difference on the factuality of the embedded predicate between having or not the negative adverb, but not between having the embedded verb in the subjunctive or in the indicative mood. However, the results from fitting different models to specific subsets of the corpus suggest that there is a small significant difference between the indicative and the subjunctive categories in specific cases.

The analysis presented here has focused more on what the data looks like and it has scrapped the surface of why it looks like that. Therefore, an important line of future work is a thorough analysis of the annotations in terms of what causes the results found. A second line of work is a different statistical analysis. The methods chosen here assume that there is one single underlying truth for each premise-hypothesis pair, but as the analysis of the disagreement patterns has shown, there is a non-neglectable number of pairs for which this is not the case. Consequently, methods that expect disagreement, like the entropy distribution seen in Nie et al. (2020), might be insightful. The third and last line of work proposed is the inclusion of out-of-sentence context in the corpus, especially since it was recommended in Manning (2006). The question about its inclusion was already raised while designing the corpus, but it was disregarded due to its cumbersome implementation and the increased difficulty in the analysis. But given the results obtained and the influence of context in mood alternation (Faulkner, 2021), adding context to the pairs could yield more informative results.

Acknowledgments

First I want to thank my Master Thesis supervisors Lucia Donatelli and Vera Demberg, thanks to their guidance I run this study. Secondly, I want to acknowledge and express my gratitude to Toloka (Pavlichenko et al., 2021), who provided the funds to collect the annotations.

References

María Amparo Alcina Caudet. 1999. *Las expresiones referenciales. Estudio semántico del sintagma nominal*. Ph.D. thesis, Universitat de València.

Rune Haubo B Christensen. 2018. Cumulative link models for ordinal regression with the r package ordinal. *Submitted in J. Stat. Software*, 35.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

JS. Dario. 2017. El quijote.

M. Davies. 2016. El corpus del español.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *LREC*.

Tris J Faulkner. 2021. *A Systematic Investigation of the Spanish Subjunctive: Mood Variation in Subjunctive Clauses*. Georgetown University.

Ana Fernández-Montraveta and Gloria Vázquez. 2014. The sensem corpus: An annotated corpus for spanish and catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, 10(2):273–288.

Ana María Fernández Montraveta, Hortènsia Curell i Gotor, Glòria Vázquez García, and Irene Castellón Masalles. 2020. The tagfact annotator and editor: A versatile tool. *Reproducció del document publicat a: Research in Corpus Linguistics, 2020, vol. 8, núm. 1, p. 131-146*.

P. Gamallo, M. García, R. Martíez-Castaño C. Piñeiro, and J.C. Pichel. 2018. Linguakit: a big data-based multilingual tool for linguistic analysis and information extraction. In *In Proceedings of The Second International Workshop on Advances in Natural Language Processing (ANLP 2018) co-located at SNAMS-2018*, pages 239–244.

Anastasia Giannakidou. 2014. (non) veridicality, evaluation, and event actualization: evidence from the subjunctive in relative clauses. In *Nonveridicality and Evaluation*, pages 17–49. Brill.

Anastasia Giannakidou and Alda Mari. 2015. Mixed (non) veridicality and mood choice with emotive verbs. In *CLS 51*.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*, volume 1. Advanced Analytics, LLC.

Kaggle. 2018. Old newspapers.

John Lyons. 1995. *Linguistic semantics: An introduction*. Cambridge University Press.

Christopher D Manning. 2006. Local textual inference: it's hard to circumscribe, but you know it when you see it–and nlp needs it.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational linguistics*, 38(2):301–333.

Errapel Mejías-Bikandi. 1998. Pragmatic presupposition and old information in the use of the subjunctive mood in spanish. *Hispania*, pages 941–948.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

N. Pavlichenko, I. Stelmakh, and D. Ustalov. 2021. Crowdspeech and voxdiy: Benchmark datasets for crowdsourced audio transcription. *arXiv preprint arXiv:2107.01091*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

RAE Real Academia Española. 2011. *Nueva gramática de la lengua española: Manual*. Espasa.

Alexis Ross and Ellie Pavlick. 2019. How well do nli models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240.

David Sánchez-Jiménez. 2011. Una aproximación teórica a la definición del modo verbal español.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227–268.

Patrick E Shrout. 1998. Measurement reliability and agreement in psychiatry. *Statistical methods in medical research*, 7(3):301–317.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Amalia Vanacore and Maria Sole Pellegrino. 2022. Robustness of κ-type coefficients for clinical agreement. *Statistics in Medicine*, 41(11):1986–2004.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.