RESLA
Revista Española de Lingüística Aplicada

https://matrix.aesla.org.es/RESLA
ISSN 0213-2028 | e-ISSN 2254-6774

Assessing factuality in Spanish: do experts and non-experts agree?

Elisabet Comelles [1]*, Hortènsia Curell [2], Juan Aparicio Mera [1] & Irene Castellón Massalles [1]

[1]Universitat de Barcelona, Spain
[2]Universitat Autònoma de Barcelona, Spain

Abstract: Event veridicality and factuality are key to natural language understanding. Deciding whether the events mentioned in a text are viewed as happening or not is one of the key features in fake news detection. That is one of the reasons why factuality has been widely studied and enormous efforts have been made to annotate corpora and develop tools that automatically identify factuality and veridicality. Most of the studies carried out use English corpora annotated by non-experts; however, little has been said about factuality in Spanish and on how appropriate the use of non-expert annotators is. Based on the discussion above, this study aims to a) shed some light on how non-expert Spanish speakers interpret the various factuality markers, b) determine if the age of the informants is a relevant factor when assessing the degree of factuality, and c) explore whether the experts and the non-experts' interpretations differ.

Keywords: *corpus, factuality markers, polarity, news, Spanish.*

Evaluación de la factualidad en español: ¿hay acuerdo entre expertos y no expertos?

Resumen: La vericidad y factualidad de un hecho o situación son vitales para la comprensión del lenguaje natural. Decidir si se considera que los hechos o situaciones de un texto ocurren o no es una de las características principales en la detección de noticias falsas. Esta es una de las razones por las que la factulidad ha sido ampliamente estudiada y se han realizado muchos esfuerzos para anotar corpus y desarrollar herramientas que detecten automáticamente la factualidad y vericidad. La mayoría de estudios realizados usan corpus en inglés anotados por informantes no expertos; sin embargo, la factualidad en español y cómo de apropiado es el uso de anotadores no expertos se ha estudiado muy poco. Teniendo en cuenta lo anterior, el objetivo de este estudio es: a) arrojar algo de luz sobre cómo hablantes de español no expertos interpretan los diferentes marcadores de factualidad; b) determinar si la edad de los informantes es un factor relevante para evaluar el grado de factualidad; y c) explorar si las interpretaciones de informantes expertos y no expertos difieren.

Palabras clave: *corpus, marcadores de factualidad, polaridad, noticias, español.*

How to cite: Comelles, E., Curell, H., Aparicio Mera, J. & Castellón Massalles, I. (2026). Assessing factuality in Spanish: do experts and non-experts agree? *Revista Española de Lingüística Aplicada, 39*(2), 228-264. https://doi.org/10.58859/resla.1545

**Corresponding author:* elicomelles@ub.edu

AESLA
Asociación Española de Lingüística Aplicada

Available under the CC BY 4.0 license. ©

1. Introduction

Factuality is a widely studied phenomenon, especially in Natural Language Processing (NLP), given that determining whether a given event is perceived as having happened or not is one of the key features in fake news detection. There is a vast body of literature about factuality in English (Saurí & Pustejovsky, 2009, 2012; De Marneffe et al., 2012; Soni et al., 2014; van Son et al., 2014; Lee et al., 2015; Prabhakaran et al., 2015; White et al., 2016; Stanovsky et al., 2017; Rudinger et al., 2018; Miller, 2020), but just a few studies about Spanish (Wonsever et al., 2016; Tracey et al., 2019; Fernández-Montraveta et al., 2020; Rosá et al., 2020; Barrios, 2022).

The study presented in this paper is part of the TAGFACT project (Alonso et al., 2018; Fernández-Montraveta et al., 2020), which studied the expression of factuality in Spanish newspaper articles. Within this project factuality is understood as the degree of commitment with which the author of a text presents the events (from now on used to include all kinds of situations) narrated, which is the approach generally used in NLP. In texts, and more specifically in journalistic texts, authors present events with varying degrees of commitment towards their certainty (from totally certain to totally uncertain), which may influence the readers' perceptions. Obviously, the author's attitude towards the truth value of the events presented in texts cannot be accessed. What can be accessed, though, is how the degree of certainty is perceived by readers. In this regard, the aims of this paper are: 1) observing how factuality markers in Spanish affect the participants' interpretation of factuality; 2) checking whether experts and non-experts agree on the assessment of factuality; and 3) determining if age difference plays a role when assessing the factuality of an event.

This paper is organised as follows: first, a literature review on the concept of factuality and factuality-related studies is provided; secondly, the methodological approach followed to carry out this study is described; thirdly, the results of both the quantitative and qualitative studies carried out are presented and discussed; finally, some conclusions on the results obtained are drawn.

2. Theoretical framework

The approach to factuality used in this paper (and in the TAGFACT project), as explained above, is the one typically followed in NLP: the degree of commitment of the author towards the truth of the event narrated. In the TAGFACT project, various factors were taken into account in order to determine the factuality of events narrated by verbs: event type, reference time, tense and polarity, together with sources beyond the author of the piece of news.

Languages have different kinds of factuality markers, at various levels, which can co-occur and interact in the same sentence. In Spanish, at the morphosyntactic level, we find verb tense forms (infinitive, past participle, present, future, present perfect, perfective past, imperfective past, subjunctive, conditional), constructions (conditional, purpose) and

verb periphrases, especially modal ones, such as *poder* 'can' + infinitive. Lexically, there are adverbs (*quizás* 'maybe'), adjectives (*posible* 'possible') and nouns (*certeza* 'certainty'), together with negative polarity expressions such as *no* 'no', *nadie* 'nobody', or *nunca* 'never'. At the discourse level, voices other than the author and evidentiality also play a role. All texts are created by an author, obviously, but other sources of information can also appear in sentences, which is quite frequent in the news. This allows the journalist to express a distance (lack of commitment) from the events narrated in the news item. As for evidentiality, specifying whether a given event has been directly witnessed or heard from someone also indicates how committed the speaker is towards what is being explained.

Several studies have been carried out to annotate and assess the factuality of texts, although most of them have focused on the annotation of factuality in English. The best known work on factuality annotation and detection is Saurí (2008), where the author presents FactBank, a factuality corpus that consists of 208 documents and 8,837 events manually annotated by two experts. In Saurí (2008), the annotators had to choose from a set of values that combined the probability of an event and its polarity. These values included six fully committed values that ranged from possibly not (PS-) to certainly (CT+), and two underspecified values (i.e., partially underspecified and fully underspecified). In addition, she also allowed for non-applicable.

Saurí's work reinforced the interest of the NLP community in this area and boosted the research on factuality. Following Saurí's study, De Marneffe et al. (2012) used a subset of 642 sentences from FactBank and recruited 10 "linguistically naive annotators" (De Marneffe et al., 2012, p. 303) located in the USA. The annotators were asked to decide if the event in bold described in the sentence did or will happen, using the set of values in FactBank. This study differs from Saurí's in that annotators were allowed to use their knowledge of the world when assessing the sentences. The results of their study show that lexical markers interact with pragmatic context, especially in the case of "absolutism".[1] Later on, Soni et al. (2014) studied how factuality is perceived in social media, namely in Twitter. They compiled a corpus of 1,170 tweets that were rated using a 5-point Likert scale ranging from "[-2] Certainly False" to "[+2] Certainly True". They allowed for "[0] Uncertain" and "Not Applicable". Ratings were gathered by means of Amazon Mechanical Turk. The authors conclude that the perceptions of factuality were influenced by the source-introducing predicates (i.e., cue), while extra-linguistic factors (e.g., source, author, etc.) did not seem to be relevant. Lee et al. (2015) performed an experiment where non-expert annotators were asked to first annotate events in a sentence, and secondly annotate factuality using a 7-point Likert scale ranging from -3 (certainly did not happen according to the author) to 3 (certainly happened according to the author). 0 meant that the author was neutral and did not express any bias towards the factuality of the event. Data were gathered using Crowd-Flower and for each example 5 judgements were collected and aggregated. The

[1] De Marneffe et al., 2012 use this term to refer to the fact that "too many lexical markers conveying certainty might in some cases undermine the speaker's credibility (e.g., in a car salesman pitch) and actually incite mistrust in the hearer/reader. In such instances, lexical markers only are not good indicators of the veridicality of the event, but the pragmatic context of the utterance needs to be taken into account to fully appreciate the interpretation people assign to it." (De Marneffe et al., 2012, p. 330).

corpus annotated consisted of 4,243 sentences from TempEval-3 (Uzzaman et al., 2013). The factuality ratings obtained were compared to FactBank annotations for overlapping events and results showed a strong correlation between the non-experts' ratings and FactBank ratings. More recently, Miller (2020) studied how 227 readers decided on the credibility and factuality of a statement about an event. The corpus used consisted of a total of approximately 4,000 events taken from a collection of witness testimony. Based on the results obtained, the author proposed a new schema to assess veridicality, with more gradations of certainty and positive evidence, and more allowances for counterfactual or uncertain statements.

Although factuality has been widely explored in English, the research on factuality in Spanish is rather scarce and can be reduced to the studies carried out by Wonsever et al. (2016), Tracey et al. (2019), Alonso et al. (2018), and most recently Tracey et al. (2022) and Barrios (2022). Wonsever et al. (2016) adopted Saurí's (2008) model with some adaptations and compiled a corpus of 875 events annotated by two experts using using a set of 6 values that indicate the occurrence (R) or non-occurrence of the event (NR), scheduled future (FP), denied future (FN), possible occurrence of the event (POS) and undefined (IND). The global inter-annotator agreement was quite high, reaching 90.4%.

Tracey et al. (2022) present the BeSt corpus, which consists of English, Chinese and Spanish data annotated for belief and sentiment. The Spanish subcorpus is composed of a total of 262 documents that come from two genres (news and forum discussions). The data were annotated by trained annotators with native fluency in Spanish. Events, the participation of an entity in an annotated event, and relations were annotated using four degrees of belief annotation: committed belief, non-committed belief, reported belief, and not applicable. Authors found that non-committed beliefs were very rare and that news had more reported belief than discussion forums. The BeSt corpus was used in 2016 and 2017 as a basis for evaluation in the DARPA DEFT program.

Most of the few studies carried out in Spanish have relied on trained annotators, showing a clear difference with those in English, in which the use of non-experts is rather common. In our study, we try to bridge this gap and explore how reliable non-experts are to annotate factuality in Spanish, as well as examine how they interpret factuality markers.

3. Methodology

In order to carry out this experiment, a survey was passed to 90 native Spanish speakers organized in two groups (G1 and G2). Group G1 consists of 60 university students under 30, whereas G2 consists of 30 informants above 30. The survey is composed of 29 simplified sentences extracted from the TAGFACT project corpus (Alonso et al., 2018), a corpus of Spanish news compiled within the framework of the TAGFACT project. The sentences were simplified (or "normalized", according to Saurí's terminology) to make syntactically complex sentences easier to understand and to reduce syntactic ambiguity, which may affect the participants' assessment of the sentences. This process was performed following

Saurí's guidelines (Saurí, 2008). The degrees used to annotate and range factuality in previous studies varies rather widely, from 9 in Saurí & Pustejovsky (2009, 2012) and 8 in Miller (2020) to 3 in Tonelli et al. (2014), Prabhakaran et al. (2015), and Colomer et al. (2016). For the purposes of this paper, a scale of 5 was chosen, because it allowed us to capture subtle differences in how factuality is perceived by readers. Thus, every sentence presented to the informant was followed by an associated question that the informant had to answer by a 5-point Likert scale, ranging from 1 [factual] to 5 [counterfactual] (see Figure 1). In addition, a distractor was added after every 5 questions in order to prevent informants from being aware of the purpose of the survey.

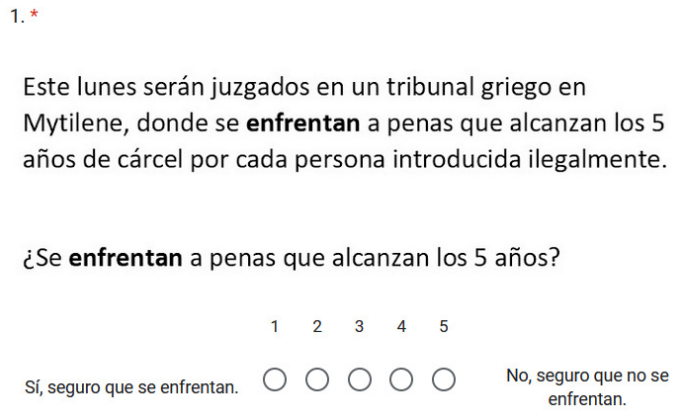


Figure 1. An example of a question in google forms.

The survey started with some brief instructions for the participants (see Figure 2), who were also asked to provide their age and proficiency in Spanish.

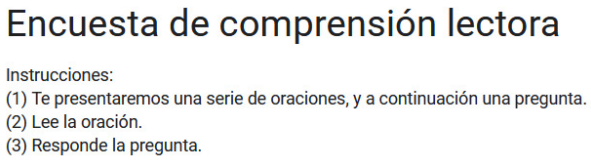


Figure 2. Survey instructions.

The following steps were taken to create the survey:

(1) Selection of the sentences for the experiment. A total of 70 sentences containing factual markers were initially extracted from the TAGFACT project. corpus. The factual markers included traits such as verb tenses (past, present, future, conditional, etc.), adverbial negation (*no* 'no', *nunca* 'never', *jamás* 'never') or negative indefinite pronouns (*nadie* 'nobody', *nada* 'nothing', *ninguno* 'none', *ninguna* 'none'), verbs (*parecer* 'seem', *intentar* 'try', *querer* 'want'), and adverbs (*seguramente* 'probably') among others.

(2) Selection of 29 candidate sentences out of the 70 initial sentences chosen. The final 29 sentences were selected to ensure the presence of a variety of polarity and commitment values, as well as a range of factuality-related traits (e.g., verb tenses, modality, and certainty markers).

(3) Simplification of the 29 sentences. The 29 selected sentences were simplified and tweaked following Saurí's guidelines (Saurí, 2008). This process consisted in:

- Highlighting the predicate (boldface) to be assessed by the informants. For each sentence, a single predicate was assessed.
- Shortening the sentences that were excessively long. Lengthy syntactically complex sentences were transformed into simple sentences. For example, sentence (a) in (1) was reduced to sentence (b).

1.[2] a. *Que su comparecencia como presidenta en la Asamblea regional no iba a dar mucho de sí, se enfundaba una camiseta de Juego de Tronos en la que se autoproclama la Khaleesi del PP y triunfaba en las webs que jamás* ***reservarían*** *un párrafo a ese debate parlamentario.*

b. *Las webs jamás* ***reservarían*** *un párrafo al debate parlamentario de la Asamblea regional.*

- Creating the question. Every sentence was followed by an associated positive question that the participants had to answer (see example (2)).

2. *A su alrededor todo es misterio y nadie quiere* ***aventurar*** *la posibilidad de que renuncie.*

¿Alguien aventura que el diputado puede renunciar?

- Avoiding certainty markers in the associated question. If the predicate included certainty markers, they were removed from the associated question. This may also involve changing the verb tense (see example (3)).

3. *Garrido fue el encargado de anunciar que el trabajo de fin de carrera de Cifuentes no se* ***haría*** *público "para no seguir estirando el chicle".*

¿El trabajo de Cifuentes se ***hará*** *público?*

- Avoiding the use of infinitives in the associated question. If the predicate assessed was an infinitive, the verb in the associated question was conjugated. Some strategies to do so involved introducing a subject, using an impersonal pronoun or a reflexive passive construction (see example (4)).

4. *"En manos del PNV. ¿En serio? ¡Vaya con la gran nación española, qué incapacidad de* ***llegar a acuerdos****!", reprochó.*

¿Se ***ha llegado*** *a acuerdos?*

[2] The English translation of all the examples has been included in Appendix 1.

- Anonymizing both sentence and question. If the sentence contained some proper names easily identifiable by the reader, they were changed for another one in order to reduce the speaker's world knowledge, which could influence the assessment of the factuality (see example (5)).

    5. *Garrido fue el encargado de anunciar que el trabajo de fin de carrera de Cifuentes no se* ***haría*** *público "para no seguir estirando el chicle"*

       ¿El trabajo de Cifuentes se ***hará público****?*

       Larson fue el encargado de anunciar que el trabajo de fin de carrera de Rosell no se ***haría público*** *"para no seguir estirando el chicle"*

       ¿El trabajo de Rosell se ***hará público****?*

The survey was developed using Google Forms and conducted in two phases: first, a total of 60 university students aged 20 to 25 were asked to complete the survey monitored by one of the researchers (G1); secondly, the survey was published in social networks and a sample of 30 individuals over 30 was chosen (G2). In addition, a hypothesis for each question was agreed upon by the researchers using the same 5-point Likert scale that the participants were asked to use. This was done in order to compare the answers provided by the participants (non-experts) to those agreed by the 4 experts. These values can be found in Appendix 1 (see hypothesis value).

Finally, both a quantitative and a qualitative analysis were conducted. The statistical analysis of the two samples (G1 and G2) was aimed at: a) checking whether there was any difference between the assessments provided by the two groups of participants (Chi-squared distribution); and b) comparing if the experts and non-experts assessments differed (Pearson Correlation Coefficient). The qualitative analysis focused on how non-experts interpret factuality markers, and examined in detail those aspects in which experts and non-experts differ.

4. Quantitative analysis

4.1 Chi-squared distribution

In terms of statistical techniques, we followed Peña (1988), Martín and Ruíz-Maya (2001), Monge and Juan (2002), and Box et al. (2008) for non-parametric statistics, and used the χ2 distribution (1).

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

This value will be the sum of non-negative numbers. The numerator of each term is the difference between the observed frequency (Oi) and the expected frequency (Ei). Therefore, the closer both values are to each other, the smaller the numerator will be, and vice versa. The denominator allows the size of the numerator to be relativized.

The previous ideas suggest that the lower the value of the statistic, the more coherent the observations obtained will be with the expected values. On the contrary, large values of this statistic will indicate a lack of agreement between the observations and what is expected. In this type of contrast, the null hypothesis is usually rejected (the observed values are consistent with the expected ones) when the statistic is higher than a certain critical value.

We applied the tests of Goodness of Fit, Homogeneity between two samples and Independence between two variables.

- The Goodness-of-Fit test consists in determining whether the data of a certain simple random sample of size n correspond to a certain theoretical distribution. In other words, this test checks if the distribution of possible answers is a random distribution (uniform distribution) (H_0) or whether the reader shows a preference for a certain value (H_1).
- The Homogeneity test checks whether the data corresponding to two or more samples are homogenous, that is, if they belong to the same population. The null hypothesis (H_0) states that they are not homogenous, whereas the alternative hypothesis (H_1) states that they are. This methodology is very useful to compare different samples and draw conclusions on the equality of their population distribution.
- The test of independence determines whether two variables referring to individuals in a population, in this case the age of individuals and the choice of answers, are related. Unlike the previous tests, this test checks the relation between two variables (i.e., age and the value of the answers). The null hypothesis (H_0) states that the two variables are dependent, whereas the alternative hypothesis (H_1) claims that they are independent.

In order to apply the Goodness-of-Fit test, we carried out two experiments. In the first one, we considered sample (G1) and, in the second, we considered sample (G2). In these cases, it is necessary that the variable under study (i.e., the choice of answers by the participants) is categorised into classes or groups (i.e., 1, 2, 3, 4 and 5 in our case).

The experiment consists in:

(1) answering 29 questions.

(2) for each question, there is a k number of possible answers (i.e., 1, 2, 3, 4, and 5).

(3) each of the k possible answers has an associated probability of occurrence p_i ($p_1 + p_2 + \ldots + p_k = 1$), which remains constant during the development of the experiment.

(4) the experiment will result in a set of observed frequencies (O_1, O_2,..., k) for each question. Obviously, $O_1 + O_2 + \ldots + O_k = n$.

(5) n = 60 in G1

(6) n = 30 in G2

The expected frequencies (Ei) for each answer must be calculated under the hypothesis of uniformity between the values. If the distribution of answers were uniform, all answers would have the same probability of selection and therefore the expected frequency would be approximately the same.

Each expected frequency (Ei) can be calculated by multiplying the total number of participants (n) by the associated probability of occurrence (pi), that is:

(G1) As there is a total of 60 participants, the expected frequency for each answer is:
$E_i = n * p_i = 60* 1/5 = \mathbf{12}$ i = 1, ..., k

(G2) As there is a total of 30 participants, the expected frequency for each answer is:
$E_i = n * p_i = 30* 1/5 = \mathbf{6}$ i = 1, ..., k

The results of the Goodness-of-Fit test lead to the following typology of answers, according to their dominant value:

- A single dominant value in both G1 and G2 [18 sentences].
  - The dominant value is 1 (factual) in 6 sentences [1, 4, 7, 17, 20, 22].
  - The dominant value is 3 in 2 sentences [6, 13].
  - The dominant value is 5 (counterfactual) in 10 sentences [3, 5, 8, 9, 12, 18, 21, 23, 24, 25].

  In all these cases there are statistically significant differences with the rest of possible answers, as shown in the distribution of sentence 17, being its dominant value 1 (see Figure 3).

  Sentence 17: *Pero no fueron los únicos. Otros dos voluntarios daneses* ***fueron arrestados*** *aquel día.*

  Question: *¿****Fueron arrestados*** *dos voluntarios más?*

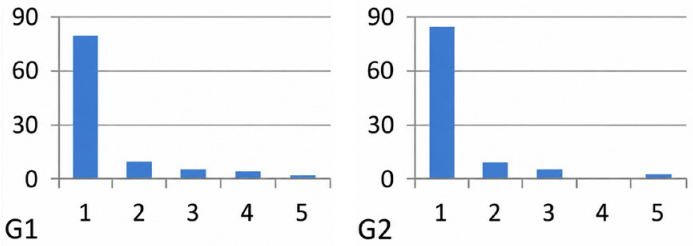


Figure 3. Distribution of sentence 17 in G1 and G2, respectively.

- 2 dominant values in both G1 and G2 [4 sentences].
  - The dominant values are 4 and 5 in two sentences [19, 28].
  - The dominant values are 3 and 2 in one sentence [11].
  - The dominant values are 5 and 3 in one sentence [14].

In all the cases above, the difference with the rest of possible answers is statistically significant, as can be seen in Figure 4, where the two possible values (4 and 5) for sentence 19 are shown.

Sentence 19: *"No voy a permitir que este tipo de cosas me* ***separen*** *de mi camino", dijo al periodista David del Cura.*

Question: *¿Este tipo de cosas le* ***separan*** *de su camino?*

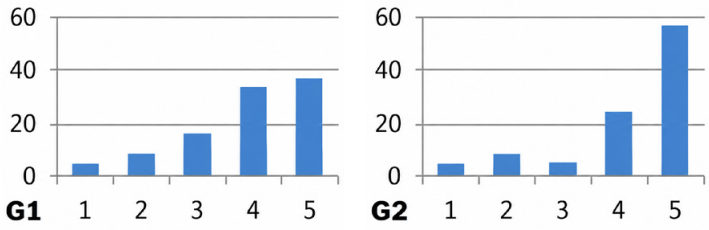


Figure 4. Distribution of sentence 19 in G1 and G2, respectively.

- 3 dominant values in both G1 and G2 [4 sentences - 10, 15, 16, 26].
  - The set of combinations varies across 3, 4, and 5; 1, 2, and 3; or 2, 3, and 5. However, in all those cases, the difference with the rest of possible answers is statistically significant. Figure 5 shows the distribution of sentence 26, being its dominant values 3, 4, and 5.

    Sentence 26: *"A los países puede no* ***gustarles*** *lo que hacemos, pero respetan lo que hacemos", ha dicho.*

    Question: *¿Les* ***gusta*** *a los países lo que hacemos?*

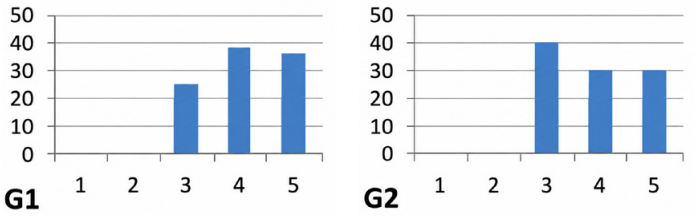


Figure 5. Distribution of sentence 26 in G1 and G2, respectively.

  - Contradictory value in both G1 and G2. In one sentence, the result in both G1 and G2 is contradictory, since the dominant values are 5 (counterfactual) and 1 (factual), showing a statistically significant difference with the rest of the answers (see Figure 6).

    Sentence 2: *En ese tiempo, nadie más* ***ha sido acusado*** *ni detenido por ese motivo en Lesbos.*

Question: *¿**Ha sido** alguien **acusado** por ese motivo?*

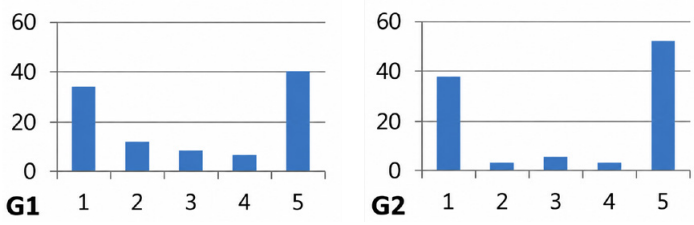


Figure 6. Distribution of sentence (2) answers in both G1 and G2, respectively.

- Random distribution of values. Two sentences show a random distribution, sentence 29 in G1 and sentence 27 in G2 (see Figure 7).

  Sentence 29: *"Han intentado **forzar** traer al presidente a rastras y de las orejas", ha llegado a decir el portavoz socialista.*

  Question: *¿**Han forzado** traer al presidente?*

  Sentence 27: *Paco, vas a volver a ganar con mayoría absoluta porque estoy seguro de que todos en Castilla-La Mancha **saben** que solo pueden depender de ti.*

  Question: *¿**Saben** todos que dependen de Paco?*

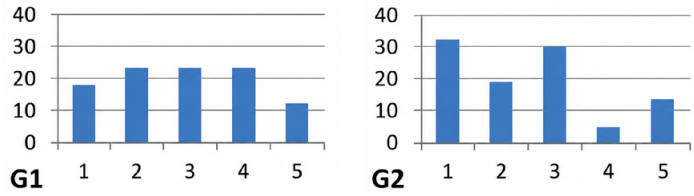


Figure 7. Random distribution of answers in sentences (29) in G1 and (27) in G2.

Once analysed the typology of answers, we can conclude that, from all the set of answers, only those corresponding to sentences 29 in G1 and 27 in G2 show a statistical test lower than the critical value for a significance level of 95%, which indicates that the answer values assigned by the informants were selected randomly. However, in the rest of sentences analysed the statistical test is higher than the critical value, thus indicating that there is a preference in the answers selected by the informants (H_1) and rejecting the null hypothesis (H_0).

Next, the Homogeneity test between the two samples was conducted. As mentioned above, this test consists of checking if several samples come from the same population. In our case, to check if group G1 (60 university students under 30) and G2 (30 participants above 30) present the same distribution in the choice of answers. For this, it is necessary

that the two measurable variables (i.e., age and the choice of answers) are represented by categories with which we will build a contingency table. Our starting point is the data obtained in the experiments we carried out for the Goodness-of-Fit test.

The distribution of answers between G1 and G2 is homogeneous in all the sentences examined, except for 3 of them, which are discussed below.

- In sentence 3, the most voted value in both G1 and G2 is 5; however, there is a remarkable difference in the percentage of informants that opted for this value in each group. In G1 this value was chosen by 55% of the informants, whereas in G2 the percentage of informants that opted for this value rises to 86%, resulting in a statistically significant difference with the rest of values (see Figure 8).

  Sentence 3: *Pero al ser acusados en grado de tentativa, porque durante su detención no iba ningún inmigrante con ellos, no* ***saben*** *realmente qué les espera.*

  Question: *¿Los acusados* ***saben*** *qué les espera?*

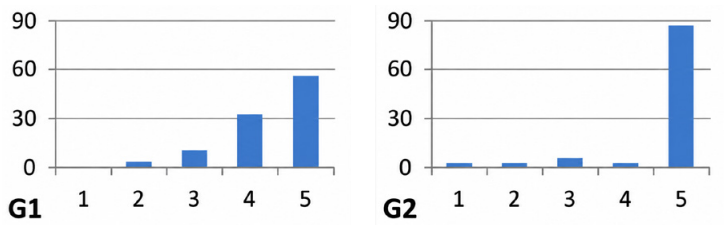


Figure 8. Distribution of answers in sentence (3) in G1 and G2, respectively.

- In sentence 13, the distribution of answers and their corresponding percentages are different in both samples. In G1 values are ranked as 3>4>2, with percentages being 46%, 33% and 11%, respectively, whereas in G2 values are ranked as 3>2>5, with percentages being 57%, 22% and 13%, respectively (see Figure 9).

  Sentence 13: *En la madrugada del mismo jueves llegó al entorno del candidato que la moción decaería por la dimisión de Montoya, y los socialistas aún no se fían de que pueda* ***dimitir*** *este viernes.*

  Question: *¿Montoya* ***dimitirá****?*

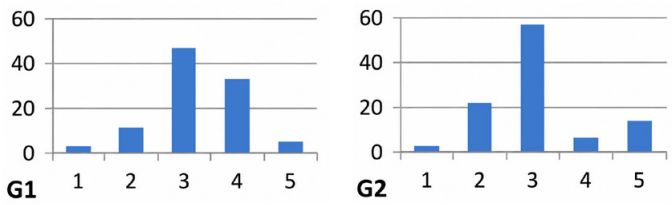


Figure 9. Distribution of answers in sentence (13) in G1 and G2, respectively.

- In sentence 29, G1 shows a random distribution of answers, while G2 presents a clear preference for values 5>3, as shown in Figure 10.

  Sentence 29: *"Han intentado **forzar** traer al presidente a rastras y de las orejas", ha llegado a decir el portavoz socialista.*

  Question: *¿**Han forzado** traer al presidente?*

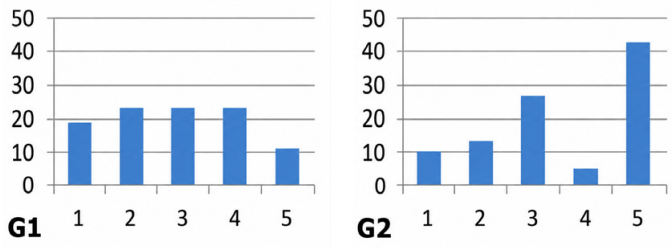


Figure 10. Distribution of answers in sentence (29) in G1 and G2, respectively.

In the three cases presented above, the chi-squared statistic is higher than the critical value for a significance level of 95%; thus, there is no homogeneity, given the statistically significant differences between the two samples. In the other 26 sentences analysed, the statistic is lower than the critical value, so the null hypothesis (H_0) is rejected and the alternative hypothesis is accepted: there is homogeneity between the two samples (H_1), that is, both samples belong to the same population.

Finally, the Independence test was conducted between two variables (i.e., the informants' age and the answer values). Our aim is to check if these variables are related to each other (we did not take into account the level of study of the participants as a variable). Although the independence test conceptually differs from the previous test (homogeneity), operationally it provides the same results.

In most of the sentences analysed the statistic is lower than the critical value, so the independence hypothesis between the two values is accepted (H_1). Only in 3 sentences (3, 13, and 29) (see Figures 8, 9, and 10), is the statistical test higher than the critical value for a significance level of 95%, so the independence hypothesis is rejected and we accept that there is a relation between age and answer values (H_0) in these three sentences. All in all, the results obtained seem to indicate that the age of the participants is independent from the answer values; in other words, age is not a factor when deciding on the factuality of an event.

4.2 Pearson Correlation Coefficient

In this section we present the results of the Pearson Correlation Coefficient (r), which was used to check the agreement between experts and non-experts. In order to calculate this correlation, the experts were asked to rate each sentence using the same 5-point Likert

scale as non-experts. In cases of disagreement, they were asked to discuss them and reach an agreed value. The experts' judgements were correlated with the non-experts' most voted value.

Correlations

		Experts	NonExperts
Experts	Pearson Correlation	1	,749**
	Sig. (2-tailed)		<,001
	N	29	29
NonExperts	Pearson Correlation	,749**	1
	Sig. (2-tailed)	<,001	
	N	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 11. Pearson Correlation Coefficient.

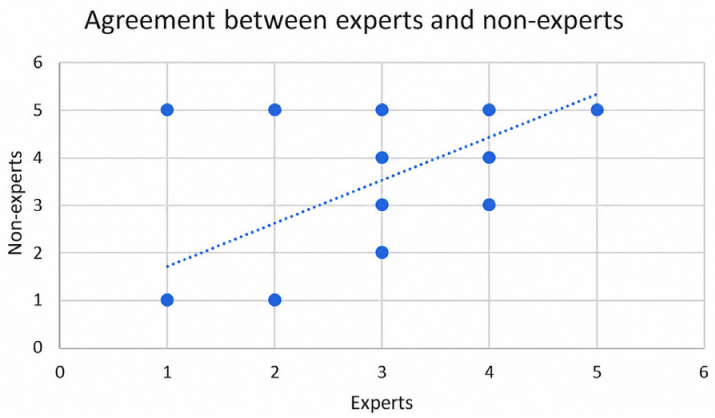


Figure 12. Agreement between experts and non-experts.

The correlation obtained between experts' and non-experts' judgements is strong (.749) and significant (see Figure 11), although there's considerable disagreement in some cases (see Figure 12).

5. Qualitative analysis

In the qualitative analysis, we first explore the relation between the answer values provided by the participants and the typology of markers. Later on, we carry out a detailed analysis of the sentences based on the agreement or disagreement between experts and non-experts, pointing out key linguistic features that trigger certain factuality values.

5.1 Participants' answers and typology of markers

The relation between the answer values provided by the participants and the various markers occurring in the sentences is shown in Table 1, where the higher the intensity of colour is, the higher the frequency of the marker. According to Table 1, the present and past tenses tend to 1 (Factual) with a different intensity, whereas the conditional and the prospective periphrasis *ir a* 'going to' + infinitive in the imperfective past tend to 5 (Counterfactual). Likewise, negative polarity shows a clear tendency towards 5 (Counterfactual), while "voice" (i.e., the presence of another speaker) is never rated as 1 (Factual).

Table 1. Relation between answer values and markers.

Markers	Answer Value				
	1	2	3	4	5
Past					
Present					
Future					
Imperfective past					
Conditional					
Infinitive					
Subjunctive					
Prospective periphrasis					
Modal verb					
Negative polarity					
Conditional construction					
Lexical verb					
ojalá 'hopefully'					
seguramente 'probably'					
Voice					

As regards multiple evidence of factuality, verb tenses combine with other markers, such as conditional *si* 'if' (e.g., sentence 6); in that sense, according to the participants' ratings, negative polarity prevails over verb tenses. Similarly, the combination of the prospective verb periphrasis (*ir a* 'going to' + infinitive) with the imperfective past results in a counterfactual interpretation (e.g., sentence 12).

> Sentence 6: *" En cualquier caso, **si** finalmente Palacín **dimite**, también es un triunfo para nosotros porque era lo que queríamos ", afirma el citado asesor.*
>
> '"In any case, if Palacín finally **resigns**, it also means a success/triumph to us because that's what we wanted", stated the aforementioned advisor.'

Sentence 12: *Lo que iba a ser una colaboración puntual acabó dando paso a una asociación, al apoyo de 10 voluntarios y a decenas de misiones.*

'What was meant to **be** a one-time collaboration eventually gave way to a partnership, the support of 10 volunteers and dozens of missions.'

As for the rest of markers, *parece* 'seems' clearly shows incompatibility with value 1 (Factual), (e.g., sentence 15), whereas s*eguramente* 'probably' is incompatible with 5 (Counterfactual) (e.g., sentence 10). On a different note, lexical verbs combine with markers such as negation and a variety of verb tenses (e.g., prospective periphrasis future, perfective past). Such lexical verbs (*querer* 'want', *creer* 'believe', *intentar* 'try', or *permitir* 'allow') are not usually placed in 1 (fact) (e.g., sentence 11).

Sentence 15: *Según el diputado catalán, parece que en la comisión **hay** barra libre.*

'According to the Catalan deputy, **there** seems to **be** a free-for-all in the Commission.'

Sentence 10: *Otros nombres que seguramente **integrarán** el Gabinete de Olivos son los de la actual portavoz en el Congreso, Josefa Céspedes y el número tres de el PSOE, Bernardo Lupo.*

'Other names that **will** probably **be part** of Olivos' Cabinet are the current spokeswoman in Congress, Josefa Céspedés and PSOE's number three, Bernardo Lupo.'

Sentence 11: *Sin embargo, en el entorno de Inocencio se cree que se puede **llegar** a acuerdos con el PNV en muchos temas, y que hay una buena sintonía inicial entre ambos partidos pese a las grandes diferencias políticas que los separan.*

'However, in Inocencio's entourage it is believed that agreements with PNV can **be reached** on other topics, and that there is initial good harmony between both parties, despite the great political differences that separate them.'

5.2 Agreement and disagreement between experts and non-experts

This section reports on a fine-grained analysis of the cases of agreement and disagreement between experts and non-experts. We will start our qualitative analysis with the twelve examples with a clear dominant value in which the informants and the researchers coincide. In six of them the dominant value is 1 (factual) and the tense and the time coincide (1 and 20 present; 4, 7, 17, and 22 past). This association between coincidence of tense and time and value 1 (factual) is so strong that even in a sentence like (7) *la organización no dejó la isla hasta* 'the organization did not leave the island until', which looks negative is correctly interpreted by all groups as positive, that is, the organization did leave the island.

In two other examples (6 and 13), the value is 3 (uncertain): (6) is a conditional sentence with *si* 'if', and in (13) a voice is present (the socialists) + *no* 'no' + *poder* 'can' in the subjunctive mood + *dimitir* 'resign'. (13) is interesting because the two groups (G1 - younger people and G2 - older people) did not behave in the same way, as shown by the homogeneity test: the second highest value in G1 is 2 (more factual) and in G2 it is 4 (less factual).

In four other sentences (3, 5, 9, and 12), the shared value is 5 (counterfactual). In (3) and (5), there is a match between tense and time (present and past respectively), together with a clear indicator of negation (*no* 'no' and *ni siquiera* 'not even'). In (3) we find again the same issue as in (13), that is, G1 shows a higher degree of factuality than G2, as signalled by the homogeneity test. In (9) there is *no* 'no' + infinitive, as a subordinate clause in the past tense referring to the past time. Curiously enough, though, some subjects assigned value 1 (factual) to this sentence, maybe because it is very long, and hence a little hard to process. In (12) there is a prospective verb periphrasis (*ir a* 'be going to'), which, in the imperfective past tends to express uncertainty. In this case, though, the uncertainty is overridden by the second part of the sentence, with a past tense expressing a past time (value 1).

The Pearson Correlation Coefficient showed that, in general, there is strong agreement between experts and non-experts; however, there are some cases with clear disagreement. We will deal now with the six examples with a single dominant value, in which the subjects and the researchers do not agree. In 5 of them (8, 18, 23, 24, and 25), the subjects' dominant score is 5 and the researchers' is 4. In three of these cases (8, 18, and 25), there is a clear negative expression (*no* 'no', *nadie* 'nobody', *jamás* 'never', and *incapacidad* 'inability'), together with other expressions that qualify the uncertainty (*no hicimos nada por* 'we didn't do anything to'; *nadie quiere aventurar la posibilidad* 'nobody wants to venture the possibility'; conditional tense). In (24) we have a quite complex sentence (*El portavoz ha deseado que ojalá toda la corrupción del PP fuera mentira* 'The spokesperson desired that he wished that all the corruption in the PP were a lie'), with a voice + desire + wish + verb in the subjunctive mood. In (21), the researchers' score is 3 and the informants' 5. Here again, there is a clear negative marker, *no* 'no', together with a verb in the conditional tense, in a clause subordinated to a saying verb. It seems to be the case that the informants grant more importance to the explicit negative markers, while the researchers pay more attention to other cues, such as verb tense.

We will continue by commenting on the four sentences with two dominant values indicating a tendency in the assessment of the factuality, although the values sometimes differ (11, 14, 19, and 28). Both groups of informants agree on the value assigned to three of them (11, 14, and 19), whereas for sentence 28 the values are swapped (see Appendix 1).

In sentence 11 (*Sin embargo, en el entorno de Inocencio se cree que se puede* ***llegar*** *a acuerdos con el PNV en muchos temas, y que hay una buena sintonía inicial entre ambos partidos pese a las grandes diferencias políticas que los separan. ¿Se* ***llega*** *a un acuerdo?*), the two dominant values assigned by both groups (3>2), as well as the experts (3), are clearly related to possibility/probability. A detailed analysis of the sentence shows

several elements that back those judgements. Firstly, although the source is explicit *en el entorno de Inocencio* 'in Inocencio's circle', the event *llegar a acuerdos* 'reach agreements' is embedded within two larger constructions *se cree que se puede* 'it is believed that agreements could be reached', which seems to diminish the effect of the explicit source. Secondly, the first predicate *se cree* 'it is believed' is a weak assertive predicate in the impersonal form, which contributes to the uncertainty value. Finally, the second predicate *se puede* 'could', also in the impersonal form, introduces an element of possibility that clearly affects the embedded infinitive *llegar* 'reach'. The combination of all the elements detailed above definitely helps convey the idea of possibility, which is also confirmed by both participants' and experts' judgements.

Both G1 and G2 also agree on the judgment of sentence 19 (*No voy a permitir que este tipo de cosas me* ***separen*** *de mi camino", dijo al periodista David Del Cura. ¿Este tipo de cosas le* ***separan*** *de su camino?*) (see Appendix 1). However, unlike the previous case, participants and experts do not agree on their evaluation. While the informants' two dominant values (5>4) show a tendency towards counterfactuality, the experts' evaluation opts for uncertainty (3). A closer look at the sentence, reveals that the event evaluated, *este tipo de cosas me separen de mi camino* 'this type of things keep me away from my path', is embedded within a bigger negative construction *no voy a permitir que* ('I won't allow that'), which in its turn is a direct quotation from the source *dijo al periodista David del Cura* ('(he) told the journalist David del Cura'). There are several characteristics that may have strongly influenced the participants' choice. First, the fact that the source is being quoted may account for the informants' trust in what the source says, hence the values heading towards counterfactuality (4>5). Secondly, the negative polarity of the predicate *no voy a permitir que* 'I won't allow that', together with the prospective verb periphrasis *ir a* 'be going to' seems to have a stronger influence than the deontic modality expressed by the verb *permitir* 'allow' itself. Although in Saurí's words "the fact that an event is allowed or required does not necessarily imply that it will ever take place in the world" (Saurí, 2008, p. 47), the combination of modality and polarity markers in this example seems to favour the participants' counterfactual interpretation of the sentence. On the other hand, the experts have opted for a more moderate interpretation, where the scope of *permitir* overtakes the role of the negative particle, thus showing a clear degree of uncertainty (3).

The role that negative particles and negative polarity play in determining the factuality of an event is also relevant in sentence 28 (*Ninguna de esas estrategias parece* ***haber funcionado****. ¿****Ha funcionado*** *alguna estrategia?*). In this case, although G1 and G2 do not coincide in their evaluation (see Appendix 1), the values assigned by both groups (G1 5>4 and G2 4>5) and the experts (4) do show a tendency towards counterfactuality. The event assessed, *haber funcionado* 'to have worked', is introduced by the predicate *parece* 'seems', which shows a slight degree of uncertainty, thus accounting for the experts' evaluation (4) as well as one of the dominant values in the evaluation performed by the participants. In addition, the polarity is negative, which matches the values 4 and 5 in the Likert scale. In this sentence, the negative polarity is expressed by the negative particle

ninguna 'none'. Similar to sentence 19, negative polarity seems to be decisive in the interpretation of the factuality of the event, especially considering the value assigned by group G1, which clearly shows a preference for the counterfactual interpretation. In that sense, the age factor seems to be relevant here, since younger participants show a more moderate tendency (see Appendix 1) that coincides with the experts' view.

The last sentence of this group, 14 (*Este jueves, Nuño pareció una dirigente **zarandeada** por la actualidad, llevada en andas por los suyos, nadie sabe hasta dónde como 'lideresa' y hasta dónde, quizá, como mártir. ¿Nuño **fue zarandeada**?*) is also assigned two dominant values, although it differs from previously analysed sentences in two aspects: first, the values provided by both groups of informants (see Appendix 1) are rather distributed along the Likert scale; and secondly, the two dominant values are not consecutive. Both groups of informants coincide in the dominant values assigned, being the most voted value 5 (counterfactual), and the second most voted value 3 (uncertain). It is also quite shocking that none of those scores agree with the experts' assessment, since the experts opt for a high degree of factuality (2), far from the counterfactuality score (5). The modality value of possibility expressed by the introducing predicate *pareció* 'seemed', explains values 3 (G1 and G2's second most voted value) and 2 (expert's judgment). The fact that the values are so distributed and distant seems to be related to the fact that *zarandeada* 'shaken' is clearly metaphorical in the statement, but can also have a literal reading in the question.

Now we will deal with several sentences that have provided results widely distributed among the different values (1-5). In some cases, there is coincidence in the distribution between the informants' and the researchers' hypotheses. In other cases, some values diverge between the two groups of informants, or they do not coincide with the hypothesis proposed by the experts.

In sentence 16 (*En la noche del domingo los analistas de los principales diarios coincidían en que solo un verdadero milagro político le permitiría a Haddad **revertir** lo que parece irreversible. ¿Haddad **revertirá** lo que era irreversible?*), all three groups (G1, G2, and researchers) agree on counterfactuality 4 (4>5>3), with a similar distribution. The variety of answers can be explained by the lexical elements that present the opinion of "the analysts"; words like *milagro* 'miracle', or expressions like *revertir lo irreversible* 'reversing the irreversible', together with a conditional tense.

An example of divergence between samples can be found in sentence 26 (*A los países puede no **gustarles** lo que hacemos, pero respetan lo que hacemos", ha dicho. ¿Les **gusta** a los países lo que hacemos?*). G1 informants propose value 4 (4>5>3) as the majority response, a value tending to counterfactuality, while in G2 the most voted value is 3 (3 >4=5), which coincides with the researchers' hypothesis (3), a value closer to uncertainty. In this sentence there are two factual markers, one of possibility (*puede* 'may') and one of negation (*no* 'no'). In G1 the negation marker is more relevant, whereas for both G2 and the researchers, the possibility marker is more influential in the determination of the factual value.

Something similar occurs in sentence 10 (*Otros nombres que seguramente **integrarán** el Gabinete de Olivos son los de la actual portavoz en el Congreso, Josefa Céspedes y el número tres de el PSOE, Bernardo Lupo.¿Josefa Céspedes y Bernardo Lupo **integrarán** el gabinete de Olivos?*). G1 and G2 coincide both in the most voted (2; almost factual), as in the other values (2>3>1). The researchers' hypothesis, on the other hand, is 3 (uncertain), due to the presence of an adverbial marker (*seguramente* 'probably') and the future tense, both of which indicate that it is not known whether the event will take place or not.

As for sentence 15 (*Según el diputado catalán, parece que en la comisión **hay** barra libre. ¿**Hay** barra libre?*), the values are highly distributed in all categories, ranging from factuality to counterfactuality. In addition, there are divergences between the samples, in G1 the majority value is 2 (2>3>5) and in G2 it is 3 (3>5>2). The researchers' hypothesis was 3, based on two pieces of evidence: the presence of a source (*el diputado catalán* 'the Catalan MP'), and of the verb *parecer* 'seem'. In this case, sample G2 coincides with the hypothesis, but not G1, which opts for a value closer to factuality. We also observe that option 4 is very poorly represented compared to 3 and 5 in both samples. Knowledge of the world may explain these differences between students and older adults: the data collection took place at a moment in which Catalan society was strongly polarised, which may have influenced the wide divergence of the answers.

Finally, we will deal with strange cases. In sentence 2 (*En ese tiempo, nadie más **ha sido acusado** ni detenido por ese motivo en Lesbos. ¿**Ha sido** alguien **acusado** por ese motivo?*), the two highest values in G1 and G2 are 4 (counterfactual) and 1 (factual). This contradiction is due to the fact the sentence and the question do not really match 'During that time, nobody else has been accused or arrested for that reason in Lesbos. Has somebody been accused for that reason?'. In the sentence, but not in the question, we have *else*, so that the question may be ambiguous. Sentence 27 (*Paco, vas a volver a ganar con mayoría absoluta porque estoy seguro de que todos en Castilla-La Mancha **saben** que solo pueden depender de ti. ¿**Saben** todos que dependen de Paco?*) is a direct quote, from an unknown speaker, in which an opinion is expressed (*estoy seguro* 'I'm sure') about a mental state of a third party (*todos saben* 'everybody knows'). The researchers' rating was 2 (quite sure that it happened), whereas G2's is random and G1s is very distributed. This might be related to the (lack of) confidence towards the utterer. Finally, sentence 29 (*"Han intentado **forzar** traer al presidente a rastras y de las orejas", ha llegado a decir el portavoz socialista. ¿Han **forzado** traer al presidente?*) is a very uncommon and convoluted sentence with two infinitives, one right after the other. G2 leans towards 5 (counterfactual) with 3 (uncertainty) as second value, while the answers in G1 are random, probably due to the syntactic complexity of the sentence together with the verb *intentar* 'try' in a past tense, which implies that we do not know whether it happened or not. The researchers' value is 3.

6. Conclusions

Factuality markers and how factuality is interpreted has been widely studied in English both from a theoretical and empirical perspective; however, the studies on factuality in Spanish are rather scarce and mainly focus on how experts understand factuality. This paper intends to help fill this gap by studying how factuality is interpreted by both experts and non-experts, exploring the type of factuality markers that favour the different readings/perceptions, and analysing if non-experts' judgements are a reliable source of information. In addition, we also examine if age is a key factor in the assessment of the factuality of an event.

The current study has followed both a quantitative and a qualitative approach. We first conducted a survey among two groups of non-experts (G1 - participants aged under 30 - and G2 - participants over 30) with a native or native-like proficiency in Spanish. The Chi-squared distribution was used to examine whether age plays a role in the interpretation of the factuality of an event. Later, the Pearson Correlation Coefficient was calculated to explore if experts and non-experts agree on their assessments of factuality. Finally, based on the results obtained by the statistical analysis, a fine-grained qualitative analysis was conducted to identify the linguistic traits that may affect the experts' and non-experts' decisions regarding factuality.

Results of the homogeneity and independence tests have shown that age does not seem to be relevant when assessing factuality. From all the sentences analysed, only in three of them was the test statistic higher than the critical value, indicating that there was a relation between age and the answers provided by the participants. On a different note, the results of the Goodness-of-Fit test have led to a classification of the sentence-answer sequences according to the dominant value provided by the participants. This typology has been the basis of the qualitative analysis of the factuality markers. This analysis has shown that those markers that are clearly factual include the past and present verb tenses, whereas those that show a tendency toward counterfactual values are negative polarity and other verb tenses such as the imperfect, the conditional, and the periphrastic future.

Regarding the comparative study between experts and non-experts, the results obtained after calculating the Pearson Correlation Coefficient show that there is a strong and significant correlation between experts and non-experts; thus, experts and non-experts tend to agree on how they interpret factuality. Interestingly, though, the qualitative analysis of those cases of disagreement has shown that informants seem to grant more importance to the explicit negative markers, while researchers pay more attention to other, more subtle cues, such as verb tenses. Finally, both quantitative and qualitative studies seem to indicate that knowledge of the world has an effect on the participants' answers, as shown by the wide divergence of answers in sentence 15.

The findings of this paper, although preliminary, may well be useful to other research studies seeking to use non-experts to annotate factuality. In addition, the classification of the factuality markers can be used to inform NLP systems aiming at automatically annotating texts with factuality values.

Acknowledgments

We would like to thank all the participants who generously devoted their time and contributed to this study. Without their willingness and cooperation, this research would not have been possible.

CRediT Author contribution / *Contribución de los autores*

Conceptualization / *Conceptualización:* Irene Castellón, Hortènsia Curell.
Formal Analysis / *Análisis formal:* Juan Aparicio, Irene Castellón, Elisabet Comelles, Hortènsia Curell.
Methodology / *Metodología:* Juan Aparicio, Irene Castellón, Elisabet Comelles, Hortènsia Curell.
Writing / *Redacción:* Juan Aparicio, Irene Castellón, Elisabet Comelles, Hortènsia Curell.
Research dataset / *Datos de investigación:* Irene Castellón, Hortènsia Curell.

Funding, data availability, and copyright / *Financiación, disponibilidad de datos y derechos de autoría*

Funding / *Financiación:* This work has been funded by the Ministry of Economy, Industry and Competitiveness - Spanish Government (TAGFACT project - FFI2017-84008-P).
Image use consent / *Consentimiento para el uso de imágenes:* Not applicable.
Conflict of interest / *Conflicto de intereses:* Non-existent.
Data availability statement/ *Declaración de disponibilidad de datos:* The sentences analysed have been extracted from the TAGFACT corpus (https://dataverse.csuc.cat/dataset.xhtml?persistentId=doi:10.34810/DATA619).
License/ *Licencia:* This article is published under the CC BY 4.0 License / *Este artículo se publica bajo la Licencia CC BY 4.0*.

Editorial history / *Fechas del proceso editorial*

Received / *Recibido:* 17/07/2023
Accepted / *Aceptado:* 05/06/2024
Published / Publicado: 15/11/2026

References

Alonso, L., Castellón, I., Curell, H., Fernández-Montraveta, A., Oliver, S., & Vázquez, G. (2018). Proyecto TAGFACT: Del texto al conocimiento. Factualidad y grados de certeza en español. *Procesamiento del Lenguaje Natural, 61*, 151–154. https://doi.org/10.26342/2018-61-20

Barrios, L. (2022). *La factualidad en las oraciones adversativas, concesivas y condicionales en español: El papel de los tiempos verbales en la anotación automática de corpus* [Doctoral Dissertation, Universitat de Lleida]. https://www.tdx.cat/handle/10803/673898#page=1

Box, G. E., Hunter, J. S., & Hunter, W. G. (2008). *Estadística para investigadores*. Editorial Reverté.

Colomer, J. P., Keyu, L., & Rambow, O. (2016). Detecting level of belief in Chinese and Spanish. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)* (pp. 22–30). https://aclanthology.org/W16-5003.pdf

de Marneffe, M. C., Manning, C. D., & Potts, C. (2012). Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics, 38*(2), 301–333. https://doi.org/10.1162/COLI_a_00097

Fernández-Montraveta, A., Vázquez, G., Curell, H., & Castellón, I. (2020). The TAGFACT annotator and editor: A versatile tool. *Research in Corpus Linguistics, 8*, 131–146. https://doi.org/10.32714/ricl.08.01.08

Lee, K., Artzi, Y., Choi, Y., & Zettlemoyer, L. (2015). Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1643–1648). https://doi.org/10.18653/v1/D15-1189

Martín, F. J., & Ruíz-Maya, L. (2001). *Estadística II: Inferencia*. Libros Aula Magna. Alfa Centauro.

Miller, B. (2020). Reading certainty: Evidence from a large study on NLP and witness testimony. *Digital Humanities 2020 Conference*. July 20–24, 2020.

Monge, J. F., & Juan, A. A. (2002). *Estadística no paramétrica: Prueba Chi-cuadrado*. Proyecto e-Math. ET-M05. Universitat Oberta de Catalunya.

Peña, D. (1988). *Estadística, modelos y métodos*. Editorial Alianza Universidad.

Prabhakaran, V., By, T., Hirschberg, J., Rambow, O., Shaikh, S., Strzalkowski, T., Tracey, J., Arrigo, M., Basu, R., Clark, M., Dalton, A., Diab, M., Guthrie, L., Prokofieva, A., Strassel, S., Werner, G., Wiebe, J., & Wilks, Y. (2015). A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)* (pp. 82–91). https://doi.org/10.18653/v1/S15-1009

Rosá, A., Alonso, L., Castellón, I., Chiruzzo, L., Curell, H., Fernández-Montraveta, A., Góngora, S., Malcouri, M., Vázquez, G., & Wonsever, D. (2020). Overview of FACT at IBERLEF 2020: Events detection and classification. In *CEUR Workshop Proceedings, 2664*, pp. 197–205. https://ceur-ws.org/Vol-2664/fact_overview.pdf

Rudinger, R., White, A. S., & Van Durme, B. (2018). Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 731–744). https://doi.org/10.18653/v1/N18-1067

Saurí, R. (2008). *A factuality profiler for eventualities in text* [Doctoral Dissertation, Brandeis University]. https://www.cs.brandeis.edu/~roser/pubs/sauriDiss_1.5.pdf

Saurí, R., & Pustejovsky, J. (2009). FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation, 43*, 227–268. https://doi.org/10.1007/s10579-009-9089-9

Saurí, R., & Pustejovsky, J. (2012). Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics, 38*(2), 261–299. https://doi.org/10.1162/COLI_a_00096

Soni, S., Mitra, T., Gilbert, E., & Eisenstein, J. (2014). Modeling factuality judgments in social media text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 415–420). https://doi.org/10.3115/v1/P14-2068

Stanovsky, G., Eckle-Kohler, J., Puzikov, Y., Dagan, I., & Gurevych, I. (2017). Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 352–357). https://doi.org/10.18653/v1/P17-2056

Tonelli, S., Sprugnoli, R., Speranza, M., & Minard, A. (2014). *NewsReader guidelines for annotation at document level. Extension of Deliverable D3*. Technical Report NWR-2014-2. Trento.

Tracey, J., Arrigo, M., & Strassel, S. (2019). *DEFT Spanish Commited Belief Annotation*. Linguistic Data Consortium LDG2019T09. Web download. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2019T09

Tracey, J., Rambow, O., Cardie, C., Dalton, A., Trang Dang, H., Diab, M., Dorr, B., Guthrie, L., Markowska, M., Muresan, S., Prabhakaran, V., Shaikh, S., & Strzalkowski, T. (2022). BeSt: The Belief and Sentiment Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 2460–2467). https://aclanthology.org/2022.lrec-1.262. https://doi.org/10.63317/4run4u9jq8mx

Uzzaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J., & Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 1–9).

van Son, C., van Erp, M., Fokkens, A., & Vossen, P. (2014). Hope and fear: Interpreting perspectives by integrating sentiment and event factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference* (pp. 3857–3864). https://doi.org/10.63317/2voodkchtwdx

White, A. S., Reisinger, D., Sakaguchi, K., Vieira, T., Zhang, S., Rudinger,R., Rawlins, K., & Van Durme, J. (2016). Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1713–1723). https://doi.org/10.18653/v1/D16-1177

Wonsever, D., Rosá, A., & Malcuori, M. (2016). Factuality annotation and learning in Spanish texts. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference* (pp. 2076–2080). https://doi.org/10.63317/3zbqbkmidsff

Appendix 1

Sentence 1 (Hypothesis Value 1)

Este lunes serán juzgados en un tribunal griego en Mytilene, donde **se enfrentan** a penas que alcanzan los 5 años de cárcel por cada persona introducida ilegalmente.

¿**Se enfrentan** a penas que alcanzan los 5 años?

*On Monday they'll be judged in a Greek court in Mytilene, where **they face** up to 5-year prison sentences for every person smuggled.*

***Do** they **face** up to 5-year prison sentences?*

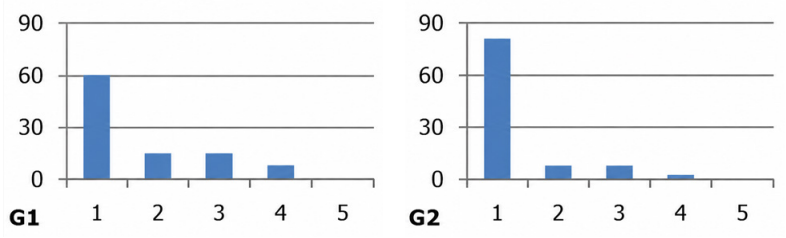


Figure 13. Distribution of answers in sentence (1).

Sentence 2 (Hypothesis Value 1)

En ese tiempo, nadie más **ha sido acusado** ni detenido por ese motivo en Lesbos.

¿**Ha sido** alguien **acusado** por ese motivo?

*In that time, nobody else **has been charged** or arrested for that reason in Lesbos.*

***Has** someone **been charged** for that reason?*

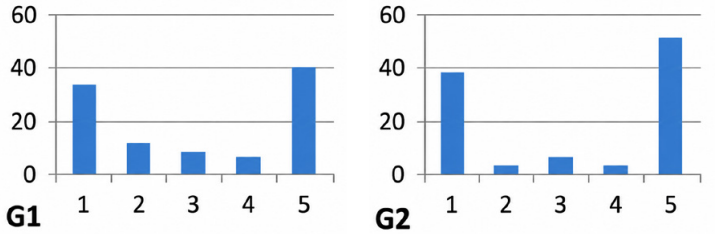


Figure 14. Distribution of answers in sentence (2).

Sentence 3 (Hypothesis Value 5)

Pero al ser acusados en grado de tentativa, porque durante su detención no iba ningún inmigrante con ellos, no **saben** realmente qué les espera.

¿Los acusados **saben** qué les espera?

*But, since they were charged with attempt because there were no immigrants with them when they were arrested, they do not really **know** what awaits them.*

***Do** the defendants **know** what awaits them?*

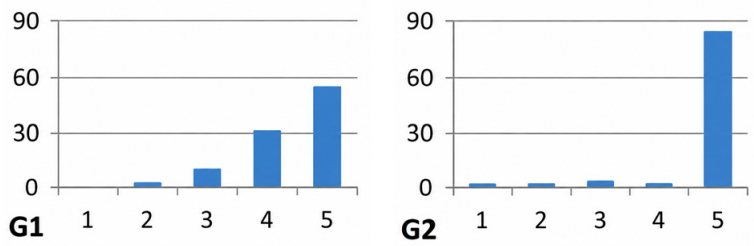


Figure 15. Distribution of answers in sentence (3).

Sentence 4 (Hypothesis Value 1)

Al no localizar el barco, volvían a puerto cuando **fueron interceptados** por los guardacostas griegos.

¿**Fueron interceptados** por los guardacostas griegos?

*Having failed to locate the ship, they were returning to port when they **were intercepted** by the Greek coast guard.*

***Were** they **intercepted** by the Greek coast guard?*

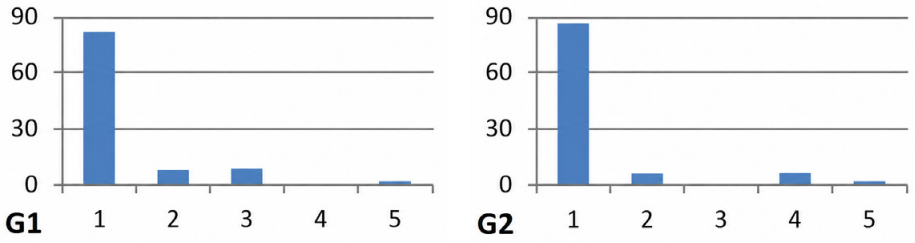


Figure 16. Distribution of answers in sentence (4).

Sentence 5 (Hypothesis Value 5)

El abogado se muestra esperanzado porque se trata de una causa "vaga e imprecisa" que carece de fundamento jurídico ya que "ni siquiera **llevábamos** a bordo a nadie esa noche", pero el hecho de que haya llegado a juicio les preocupa.

¿**Llevaban** a alguien a bordo esa noche?

*The lawyer is hopeful because it is a "vague and imprecise" case that doesn't have any legal basis since "we **did** not even **have** anybody on board that night". However, the fact that it has been taken to court worries them.*

***Did** they **have** anybody on board that night?*

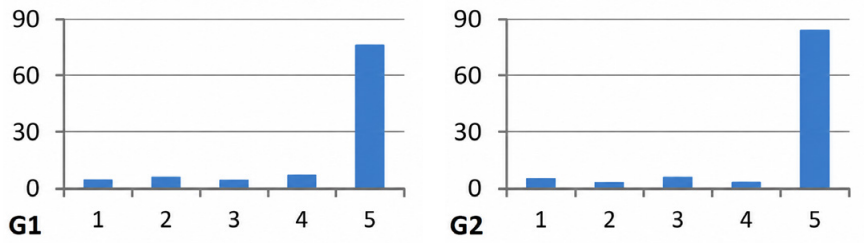


Figure 17. Distribution of answers in sentence (5).

Sentence 6 (Hypothesis Value 3)

"En cualquier caso, si finalmente Palacín **dimite**, también es un triunfo para nosotros porque era lo que queríamos", afirma el citado asesor.

¿Palacín **dimite**?

*"In any case, if Palacín finally **resigns**, it also means a success for us because that's what we wanted", states the aforementioned advisor.*

***Does** Palacín **resign**?*

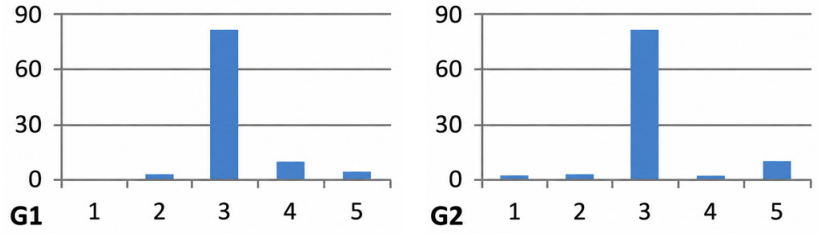


Figure 18. Distribution of answers in sentence (6).

Sentence 7 (Hypothesis Value 1)

Es más, la organización no **dejó** la isla hasta 2017.

¿La organización **dejó** la isla en 2017?

*Moreover, the organization **did** not **leave** the island until 2017.*

***Did** the organization **leave** the island in 2017?*

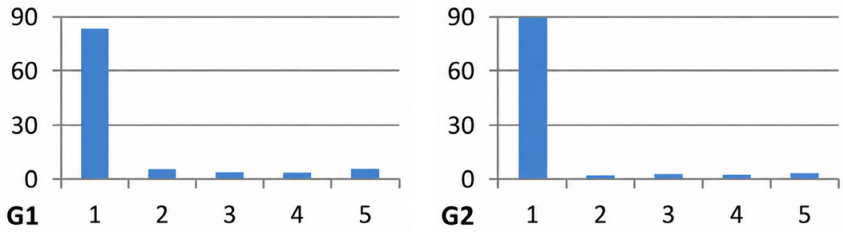


Figure 19. Distribution of answers in sentence (7).

Sentence 8 (Hypothesis Value 4)

"Pero", dice Stone, "no hicimos nada por **seguir** les el juego y porque yo, al contrario de los demócratas, no he hecho nada malo".

¿Les **siguió** Stone el juego?

*"But", Stone says, "we did not do anything **to play along** and because, unlike the Democrats, I have done nothing wrong."*

***Did** Stone **play along**?*

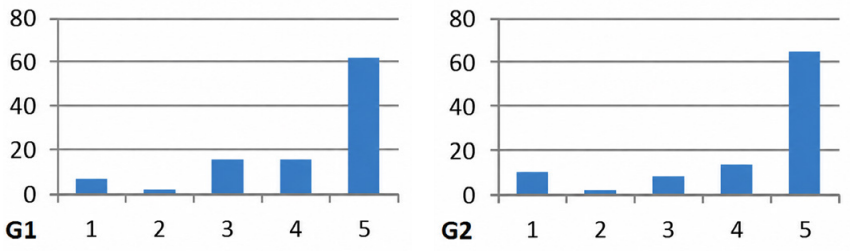


Figure 20. Distribution of answers in sentence (8).

Sentence 9 (Hypothesis Value 5)

Al no **localizar** el barco, volvían a puerto cuando fueron interceptados por los guardacostas griegos, detenidos junto a dos voluntarios daneses y acusados de facilitar el tráfico ilegal de inmigrantes e incluso de posesión de armas.

¿El barco **fue localizado**?

*Having failed to **locate** the ship, they were going back to port when they were intercepted by the Greek coast guard, arrested along with two Danish volunteers and charged with facilitating illegal smuggling of immigrants and even possession of weapons.*

***Was** the ship **located**?*

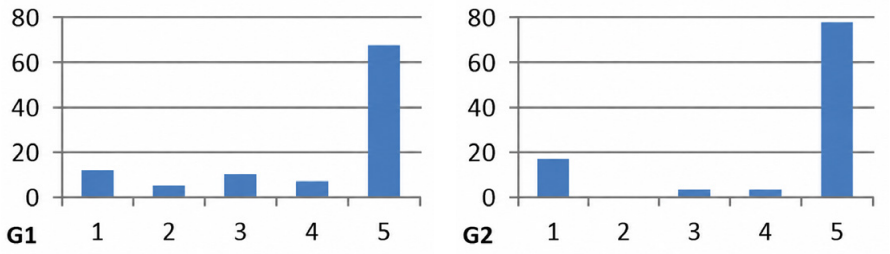


Figure 21. Distribution of answers in sentence (9).

Sentence 10 (Hypothesis Value 3)

Otros nombres que seguramente **integrarán** el Gabinete de Olivos son los de la actual portavoz en el Congreso, Josefa Céspedes y el número tres del PSOE, Bernardo Lupo.

¿Josefa Céspedes y Bernardo Lupo **integrarán** el gabinete de Olivos?

*Other names that **will** surely **be part** of Olivos' Cabinet are the current spokeswoman in Congress, Josefa Céspedés and PSOE's number three, Bernardo Lupo.*

***Will** Josefa Céspedes and Bernardo Lupo **be part** of Olivos' Cabinet?*

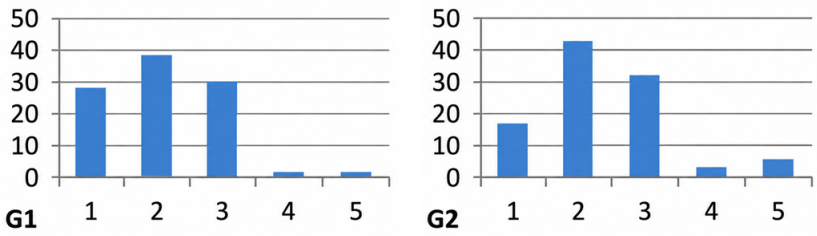


Figure 22. Distribution of answers in sentence (10).

Sentence 11 (Hypothesis Value 3)

Sin embargo, en el entorno de Inocencio se cree que se puede **llegar** a acuerdos con el PNV en muchos temas, y que hay una buena sintonía inicial entre ambos partidos pese a las grandes diferencias políticas que los separan.

¿Se **llega** a un acuerdo?

*However, in Inocencio's entourage it is believed that agreements with PNV can **be reached** on other topics, and that there is initial good harmony between both parties, despite the great political differences that separate them.*

***Is** an agreement **reached**?*

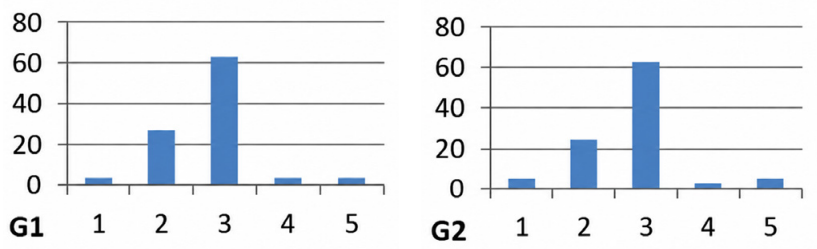


Figure 23. Distribution of answers in sentence (11).

Sentence 12 (Hypothesis Value 5)

Lo que iba a **ser** una colaboración puntual acabó dando paso a una asociación, al apoyo de 10 voluntarios y a decenas de misiones.

¿**Fue** una colaboración puntual?

*What was meant to **be** a one-time collaboration eventually gave way to a partnership, to the support of 10 volunteers and dozens of missions.*

***Was** it a one-time collaboration?*

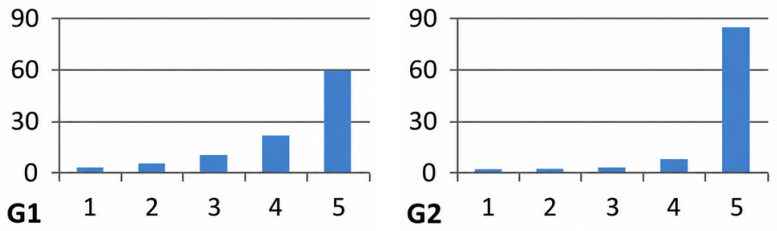


Figure 24. Distribution of answers in sentence (12).

Sentence 13 (Hypothesis Value 3)

En la madrugada del mismo jueves llegó al entorno del candidato que la moción decaería por la dimisión de Montoya, y los socialistas aún no se fían de que pueda **dimitir** este viernes.

¿Montoya **dimitirá**?

*In the early hours of Thursday morning, it was reported to the candidate's entourage that the motion would fall because of Montoya's resignation, and the socialists are still not confident that he can **resign** on Friday.*

***Will** Montoya **resign**?*

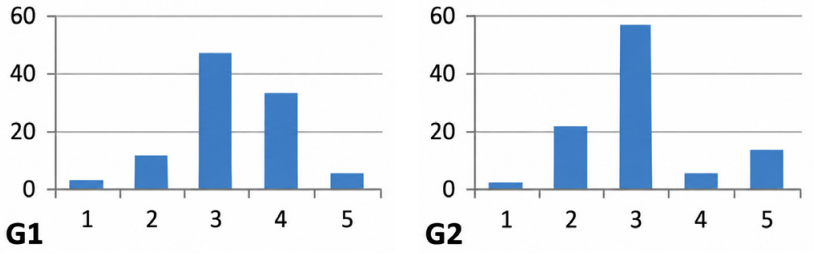


Figure 25. Distribution of answers in sentence (13).

Sentence 14 (Hypothesis Value 2)

Este jueves, Nuño pareció una dirigente **zarandeada** por la actualidad, llevada en andas por los suyos, nadie sabe hasta dónde como 'lideresa' y hasta dónde, quizá, como mártir.

¿Nuño **fue zarandeada**?

*On Thursday, Nuño seemed a leader **shaken** by the news, carried aloft by her supporters, nobody knows up to what extent as a 'leaderess" and up to what extent, maybe as a martyr.*

***Was** Nuño **shaken**?*

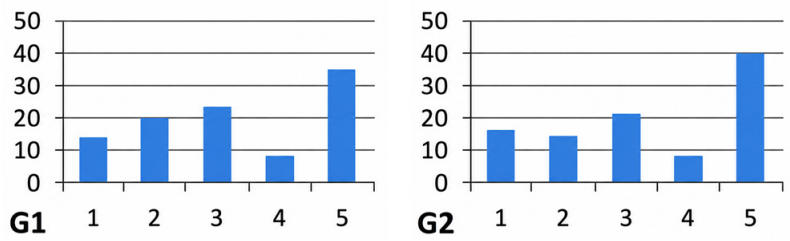


Figure 26. Distribution of answers in sentence (14).

Sentence 15 (Hypothesis Value 3)

Según el diputado catalán, parece que en la comisión **hay** barra libre.

¿**Hay** barra libre?

*According to the Catalan deputy, **there** seems to **be** a free-for-all in the Commission.*

***Is there** a free-for-all?*

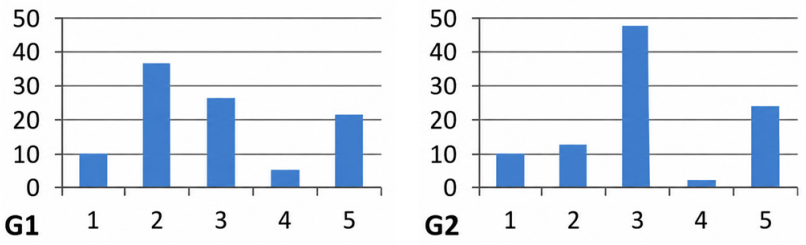


Figure 27. Distribution of answers in sentence (15).

Sentence 16 (Hypothesis Value 4)

En la noche del domingo los analistas de los principales diarios coincidían en que solo un verdadero milagro político le permitiría a Haddad **revertir** lo que parece irreversible.

¿Haddad **revertirá** lo que era irreversible?

*On Sunday evening, analysts from the main newspapers agreed that only a true political miracle would allow Haddad to **reverse** what seems irreversible.*

***Will** Haddad **reverse** what was irreversible?*

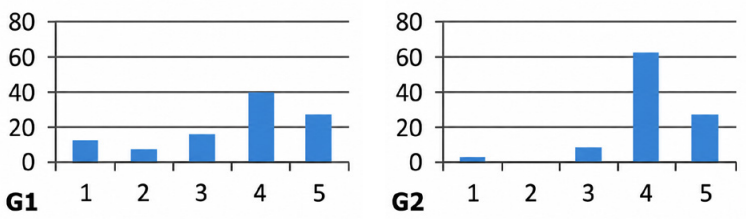


Figure 28. Distribution of answers in sentence (16).

Sentence 17 (Hypothesis Value 1)

Pero no fueron los únicos. Otros dos voluntarios daneses **fueron arrestados** aquel día.

¿**Fueron arrestados** dos voluntarios más?

*But they were not the only ones. Two other Danish volunteers **were arrested** that day.*

***Were** two other volunteers **arrested**?*

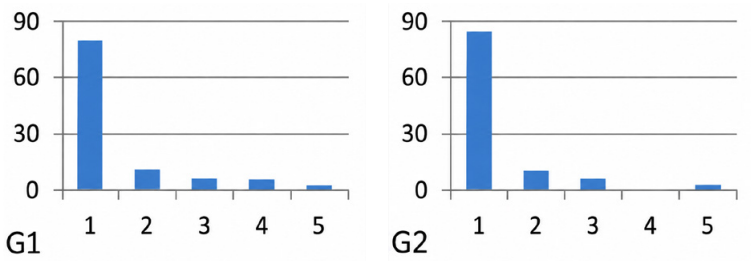


Figure 29. Distribution of answers in sentence (17).

Sentence 18 (Hypothesis Value 4)

A su alrededor todo es misterio y nadie quiere **aventurar** la posibilidad de que renuncie.

¿Alguien **aventura** que el diputado puede renunciar?

*Everything around him is a mystery and no one wants to **venture** the possibility of his resignation.*

***Does** anyone **venture** that the deputy may resign?*

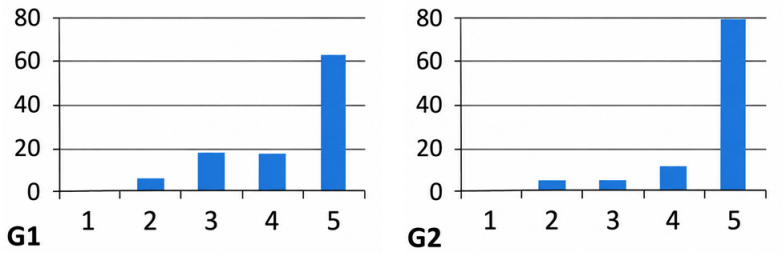


Figure 30. Distribution of answers in sentence (18).

Sentence 19 (Hypothesis Value 3)

"No voy a permitir que este tipo de cosas me **separen** de mi camino", dijo al periodista David Del Cura.

¿Este tipo de cosas le **separan** de su camino?

*"I'm not going to allow these sorts of things to **keep** me from my path", said to the journalist David Del Cura.*

*Do these sorts of things **keep** him from his path?*

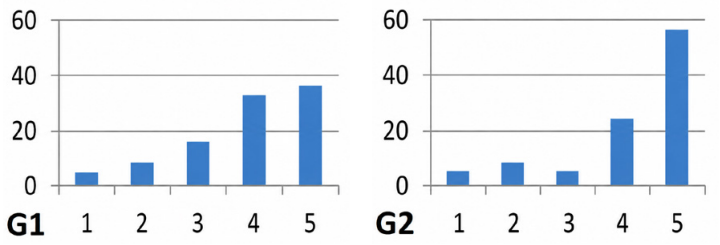


Figure 31. Distribution of answers in sentence (19).

Sentence 20 (Hypothesis Value 1)

La operación de este martes no guarda relación con estas dos investigaciones, que **lleva a cabo** la Audiencia Nacional.

¿Esta operación la **lleva a cabo** la Audiencia Nacional?

*Tuesday's operation does not have any relation with these two inquiries, **carried out** by the National High Court.*

***Is** this operation **carried out** by the National High Court?*

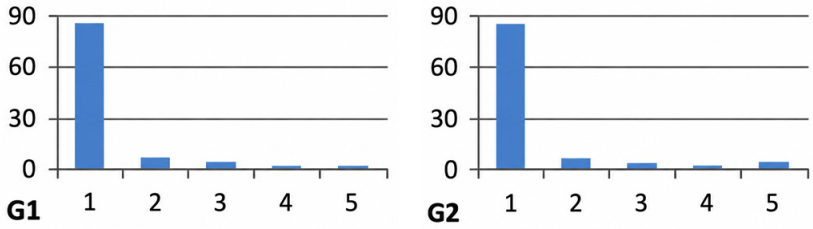


Figure 32. Distribution of answers in sentence (20).

Sentence 21 (Hypothesis Value 3)

Larson fue el encargado de anunciar que el trabajo de fin de carrera de Rosell no se **haría público** "para no seguir estirando el chicle"

¿El trabajo de Rosell se **hará público**?

Larson was in charge of announcing that Rosell's end of degree paper would not be made public "to avoid stretching it out".

***Will** Rosell's paper **be made** public?*

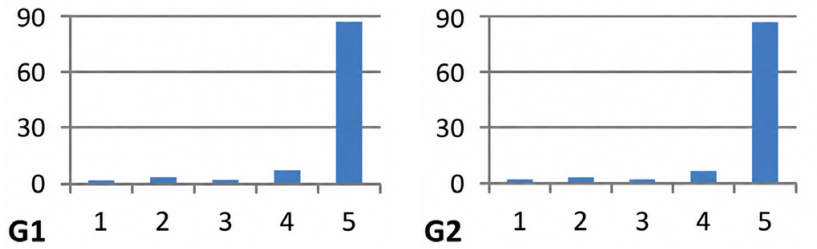


Figure 33. Distribution of answers in sentence (21).

Sentence 22 (Hypothesis Value 1)

En Onda Cero Rosell **atribuyó** la publicación de eldiario@es "a un ataque sin fundamento construido sobre la falsedad".

¿Rosell **atribuyó** la publicación a un ataque?

In Onda Cero, Rosell attributed the publication by eldiario@es "to an unfounded attack built on falsehood".

***Did** Rosell **attribute** the publication to an attack?*

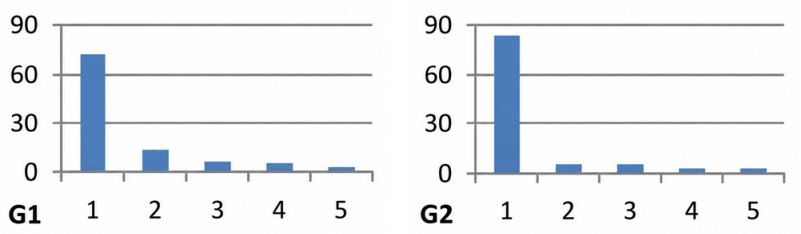


Figure 34. Distribution of answers in sentence (22).

Sentence 23 (Hypothesis Value 4)

Las webs jamás **reservarían** un párrafo al debate parlamentario de la Asamblea regional.

¿Las webs **reservarán** espacio a este debate parlamentario?

*The websites **would** never **save** a paragraph for the parliamentary debate at the regional Assembly.*

***Will** the websites **save** some space for that parliamentary debate?*

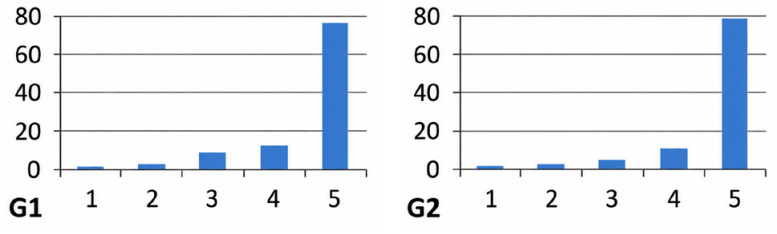


Figure 35. Distribution of answers in sentence (23).

Sentence 24 (Hypothesis Value 4)

El portavoz ha deseado que "ojalá" toda la corrupción del PP **fuera** mentira.

¿**Es** mentira toda la corrupción del PP?

The spokeperson wished all the PP corruption ***was*** *a lie.*

Is *all the PP corruption a lie?*

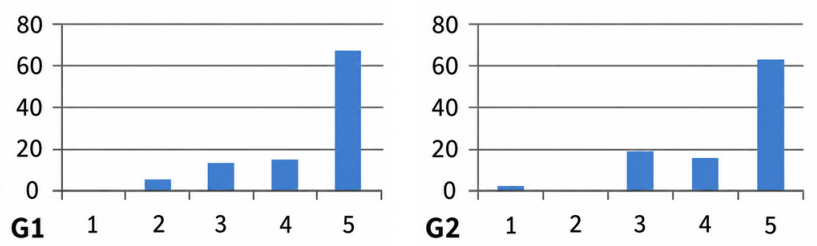


Figure 36. Distribution of answers in sentence (24).

Sentence 25 (Hypothesis Value 4)

“En manos del partido. ¿En serio? ¡Vaya con la gran nación española, qué incapacidad de **llegar a acuerdos**!”, reprochó.

¿Se **ha llegado a acuerdos**?

“In the party’s hands. Seriously? Look at the great Spanish nation, totally incapable of ***reaching*** *agreements,” he/she criticized.*

Have *agreements* ***been reached****?*

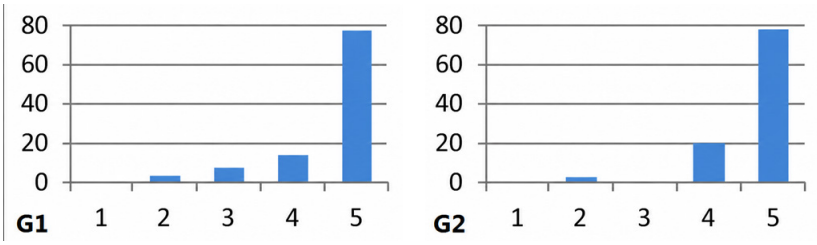


Figure 37. Distribution of answers in sentence (25).

Sentence 26 (Hypothesis Value 3)

“A los países puede no **gustarles** lo que hacemos, pero respetan lo que hacemos”, ha dicho

¿Les **gusta** a los países lo que hacemos?

“Countries may not ***like*** *what we do, but they respect what we do,” he/she said.*

Do *countries* ***like*** *what we do?*

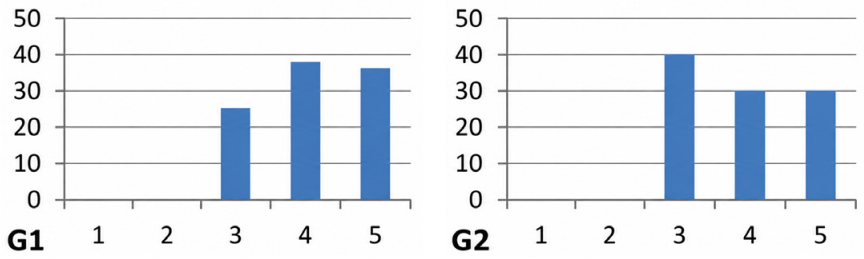


Figure 38. Distribution of answers in sentence (26).

Sentence 27 (Hypothesis Value 2)

Paco, vas a volver a ganar con mayoría absoluta porque estoy seguro de que todos en Castilla-La Mancha **saben** que solo pueden depender de ti.

¿**Saben** todos que dependen de Paco?

*Paco, you will win again by absolute majority, because I'm sure that everybody in Castilla-La Mancha **knows** that they can only depend on you.*

***Does** everybody **know** that they only depend on Paco?*

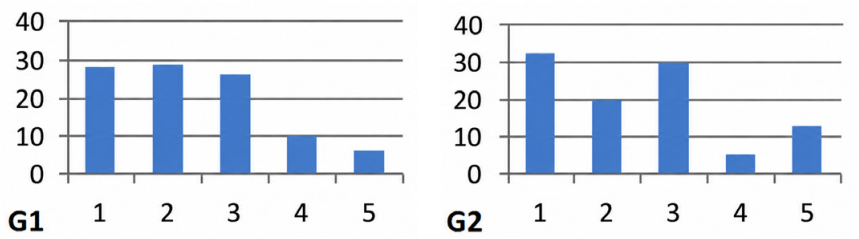


Figure 39. Distribution of answers in sentence (27).

Sentence 28 (Hypothesis Value 4)

Ninguna de esas estrategias parece **haber funcionado**.

¿**Ha funcionado** alguna estrategia?

*None of those strategies seemed to **have worked**.*

***Has** any strategy **worked**?*

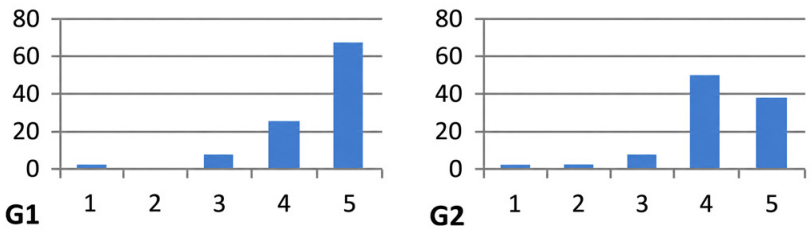


Figure 40. Distribution of answers in sentence (28).

Sentence 29 (Hypothesis Value 4)

"Han intentado **forzar** traer al presidente a rastras y de las orejas", ha llegado a decir el portavoz socialista.

¿**Han forzado** traer al presidente?

*"They have tried to **force** the president to be dragged here," the Socialist spokesman went so far as to say.*

***Have** they **forced** the president to be dragged?*

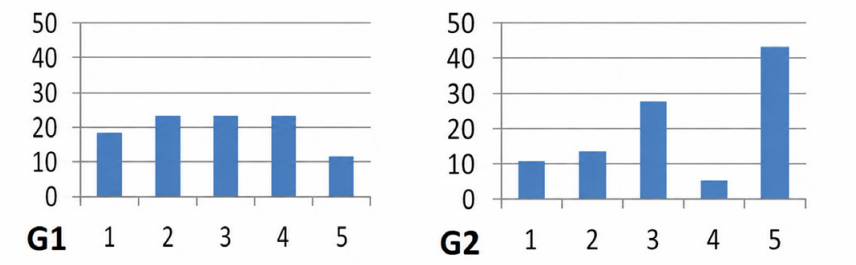


Figure 41. Distribution of answers in sentence (29).