REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2023 PUCV, Chile ⇝ 56(111) 59-75

Inter-Annotator Agreement for the Factual Status of Predicates in the TAGFACT Corpus

Estudio del acuerdo entre anotadores del estatus factual de los predicados en el corpus TAGFACT

Ana Fernández-Montraveta
UNIVERSITAT AUTÒNOMA DE BARCELONA
ESPAÑA
ana.fernandez@uab.cat

Irene Castellón
UNIVERSITAT DE BARCELONA
ESPAÑA
icastellon@ub.edu

Recibido: 12-V-2020 / **Aceptado:** 30-IX-2022
DOI: 10.4067/S0718-09342023000100059

Abstract

This paper reports on a study of the inter-annotation agreement to assess the manual annotation of the TAGFACT Gold Standard corpus. This corpus has been created as part of a larger project (TAGFACT), whose final objective is to automatize the classification of the factual status of events in a corpus of Spanish journalistic texts. In our study, six annotators labeled a corpus using the four levels of linguistic description proposed in our project to extract factual information. Each one of these levels has been assessed independently. As expected, the more fine-grained the classification is, the more problematic the annotation. This study identifies some of the most important differences and discusses the main problems encountered to obtain full agreement. We use Cohen's Kappa to measure inter-annotation agreement as well as descriptive statistical analysis.

Keywords: Factuality, inter-annotator agreement, Cohen's Kappa, journalistic texts, Spanish.

Resumen

En este trabajo presentamos un estudio sobre el acuerdo entre anotadores alcanzado en la fase de anotación manual del Gold Standard del corpus TAGFACT. Este corpus ha sido creado dentro del proyecto TAGFACT, cuyo objetivo final es automatizar la clasificación factual de los eventos narrados en textos periodísticos escritos en español. En nuestro estudio, seis anotadores han etiquetado un corpus con los cuatro niveles de descripción lingüística propuestos en nuestro proyecto para extraer información factual. Cada uno de estos niveles se ha evaluado de forma independiente. Como cabe esperar, cuanto más precisa sea la descripción lingüística, más problemática resulta la clasificación. Este estudio identifica algunas de las diferencias más importantes y se presentan los problemas que justifican que no se dé un acuerdo completo. Para el análisis del acuerdo en la anotación hemos utilizado la Kappa de Cohen, así como un análisis estadístico descriptivo.

Palabras Clave: Factualidad, acuerdo entre-anotadores, Kappa de Cohen, registro periodístico, español.

INTRODUCTION

Factuality is often described as the certainty with which the author of a message expresses their stance towards the events or situations being narrated (Saurí, 2008; Saurí & Pustejovsky, 2009). The labeling of this semantic information presents serious difficulties since it is generally a matter of interpretation, which may vary even among expert linguists.

In the field of natural language processing (NLP), the annotation of this semantic category, factuality, has experienced an increasing importance with respect to both the creation of annotated corpora with this type of information (Minard, Speranza & Caselli, 2016; Santana, Nieuwenhuijsen, Spooren & Sanders 2017; Vigus, Van Gysel & Croft, 2019) and the creation of methodologies to automate the process (Wonsever, Malcuori & Rosá, 2008; Saurí & Pustejovsky, 2012; Marneffe, Manning & Potts, 2012; Tianxiong, Peifeng & Qiaoming, 2018; Hasanain, Suwaileh, Elsayed, Barrón-Cedeño & Nakov, 2019). The relevance of this area of research is also explained by its multiple applications in other processes, such as information retrieval (Wiebe & Riloff, 2011), fact checking (Leblay, 2017), Q/A systems (Bian, Liu, Agichtein & Zha, 2008), or sentiment analysis (Matsuyoshi, Eguchi, Sao, Murakami, Inui & Matsumoto, 2010).

Currently, there are several projects (Diab, Levin, Mitamura, Rambow, Prabhakaran & Guo, 2009; Soni, Mitra, Gilbert & Eisenstein, 2014; van Son, van Erp, Fokkens & Vossen, 2014; Lee, Artzi, Choi & Zettlemoyer, 2015, Sahu & Majumdar, 2017, among others) dealing with the annotation of factuality and most of them are based, either completely or partially, on FactBank (Saurí & Pustejovsky, 2009). The above-mentioned projects deal with the annotation of English texts. For other languages, this kind of annotation is also used. This is the case of Spanish (Wonsever et al., 2008; Wonsever, Rosá & Malcuori, 2016), Italian (Minard et al., 2016; Matsuyoshi et al., 2010), Chinese (Tianxiong et al., 2018), and Japanese (Narita, Mizuno & Inui, 2013).

There are two features Factbank shares with all these projects. The basic unit of analysis is the predicate and, therefore, each predicate is analyzed individually, and events are assessed according to what is narrated in the text. That is, the certainty value assigned is determined by how events are presented in the text. A project that departs from Factbank tenants is Marneffe et al. (2012) since it links the description of events to knowledge of the world instead of just the author's stance.

Our project also takes into account the degree of certainty with which a predicate is presented. It also follows Factbank in that predicates are the units of analysis. The final goal of the annotation process is to classify predicates, at clause level, into factual classes. To perform this classification, each predicate is characterized with a value for each of the four labels of linguistic description considered: polarity, time, commitment, and event type. From the interaction between them, predicates are classified as facts, non-facts, and counterfacts at a later stage.

One of the first milestones in our project has been the creation and manual annotation of a Gold Standard (henceforth GS), presented in Section 2.1. In order to proceed with the manual annotation of the GS, an experiment was carried out to check how clear the annotation guidelines were and if they had been uniformly understood and coherently applied. This experiment also helps to see how reproducible the annotation task is and, therefore, serves a double purpose: to validate the process of manual annotation and to guarantee the reproducibility of the classification.

For this experiment, we had six annotators labeling the same training sentences for the purposes already mentioned. All of them were expert linguists who were part of the research team and willingly volunteered to participate in the experiment. In order to compute the inter-observer agreement, we decided to calculate Cohen's Kappa between each pair of annotators. This article describes the process of evaluation of the annotation, the agreement score, and the Cohen's Kappa value; moreover, it discusses the major areas of disagreement observed. In what follows, we present the project (Section 2) providing details of the GS (Section 2.1) and the annotation scheme (Section 2.2). Section 3 presents a summary of the evaluation of the agreement and, finally, Section 4 a discussion of the most relevant data.

1. The TAGFACT project

The main goal of the TAGFACT project is the creation of an automatic tool to identify the factual status of situations, that is, the degree of certainty with which situations are presented in journalistic texts. The tool will be based on the linguistic information, either implicit or explicit, present in a piece of news (triggers). Triggers will be used in the implementation of the rules of the automatic annotator; they are similar to the so-called 'modalizers'or modalization marks in discourse analysis. In our case, we deal with two types of modalizers: those marking the degree of certainty and those marking the sentence modality. The former are used to mark conviction or doubt (which are equivalent to our labels 'commitment' and 'non-commitment', respectively) while the latter are used to decide if a sentence is a candidate for further annotation (Applies or NA- Does not apply).

There are two phases in our project. In the first, a part of the corpus, the GS, has been manually annotated. In the second phase, all the knowledge collected in phase 1 is used to implement the automatic annotation tool. At the present stage, both ensuring the reliability of the manual annotation as well as the accuracy and adequacy of the annotation scheme are critical to the automation process.

The labels used in this project are partially based on existing tag sets (Section 2.2). Besides the four linguistic categories previously mentioned, the source of the information is also identified. Unless otherwise specified, the narrator by default is always the writer of the piece of news, but due to the very nature of the journalistic genre, quite often other sources are also stated in the text (either by direct or indirect speech). Taking into account polyphony is important since it provides information about possible contradictions in the narration of events in a text or provides an opportunity to the author to distance from the truth value of the narrated facts.

As mentioned above, at the most basic level of the annotation process, predicates are annotated at clause level. Besides, each predicate within a clause is treated independently. In our project, the automatic annotation focuses on the analysis of each sentence independently and, at this time, we have not worked yet on the possible relationships between the facts narrated in different sentences or between facts narrated in different texts.

The labels assigned in this step can be later modified in the case of subordinate clauses and non-finite clauses. That is, in a second step, the predicate of one of these types of clauses can be re-annotated, if necessary, depending on the main verb (by means of inheritance rules). For example, factive predicates, such as *lamentar* -regret- or *negar* -refuse- are typical examples of this type of modification. In the case of *lamentar*, the subordinate will always be a fact; in the case of *negar*, the non-finite clause will be a counterfact.

(1) Además, la vicepresidenta ha hecho del conflicto un ataque feminista y ha lamentado que el resto de diputadas no la hayan defendido por ser del Partido Popular.[1]
'In addition, the vice-president has made the conflict a feminist attack and has regretted that the rest of the female representatives have not defended her because she belongs to the Popular Party.'

In (1), the subordinate clause would not be further annotated since it is in the subjunctive mood and therefore could be interpreted as describing an unreal situation, but since it depends on the verb *lamentar*, a factive verb, it is re-annotated as a fact (commitment, positive polarity, past and event).

(2) El Tribunal Supremo de Estados Unidos falló hoy a favor de un pastelero que se negó a diseñar una tarta de bodas para una pareja homosexual alegando motivos religiosos.[2]

'The U.S. Supreme Court today ruled in favor of a baker who refused to design a wedding cake for a gay couple on religious grounds.

In (2), the non-finite form *diseñar* (design) is annotated as a counterfact (commitment, negative polarity, past and event) because it depends on the verb *negar* (refuse) in the past, which would be interpreted as introducing a counterfact. That is, *se negó a diseñar* means that he did not do it (unless this information is contradicted somewhere else in the text).

Finally, another important feature of our project is that the annotation is linked to the world described in the text. Ideologies are present in the written press. Precisely for this reason, we decided to annotate the description of the same referential piece of news from three different perspectives (ideological standpoints) so that, in the future, the commitment of the writers can be compared for the extraction of 'real' events.

1.1. Gold Standard

The GS was created with a three-fold objective: firstly, to help in the detection of the linguistic cues that trigger the factual status of a situation; secondly, to be used to check any problems the annotation scheme might pose and, finally, to be used as the benchmark for the evaluation of the automatic annotation.

The GS contains 12,475 tokens grouped into approximately 350 sentences containing a total of 1,188 predicates. As already mentioned, a part of the GS was used for the present study, the agreement corpus (AC). This AC is made up of 3 pieces of news containing 63 sentences, 280 predicates, and 1,142 tokens. It was manually annotated by 6 expert linguists, members of the research team, in order to carry out the present study.

Prior to the annotation process, texts were parsed using Freeling (Padró & Stanilovsky, 2012) and predicates automatically identified. Freeling was chosen over other parsers since it also offers a semantic graph and resolves co-referencing, which was convenient for this project that also aims to co-index events in a text (Section 2.1). In addition, it provides a good quality morphological and syntactic analysis. Lloberes, Castellón, and Padró (2015) evaluated its performance and obtained an optimal index for syntactic parsing.

1.2. Annotation scheme

In what follows, we present the labels considered in the TAGFACT annotation scheme (Vázquez & Fernández-Montraveta, 2020).

The first decision annotators faced is whether a predicate is a candidate to be annotated (Applies) with a factual value or not (NA- Does not apply). That is, clauses expressing orders, wishes, or desires; for example, are tagged as NA. For the other cases, the label chosen is Applies, which means it will be further annotated with the rest of the tags in the scheme (tagset). Four aspects of the semantics of the predicates (and sentences) annotated are considered: event types, the narrator's commitment, polarity, and time.

- **Event types**: following Vendler (1957), at a higher level we distinguish between dynamic and non-dynamic situations.
  1) The tags, Event and Mental, both describe dynamic situations that progress over time. Event includes both events and processes whereas Mental describes cognitive processes.
  2) Non-dynamic situations cover those statements expressing a property. States are further classified into Property (a state that relates a property to an entity), Abs-truth Property (sentences expressing absolute truths, that is, scientific facts or cultural beliefs) and Event Property (situations that describe eventive properties in that they refer to repeated actions or properties of events), following Vázquez and Fernández-Montraveta (2020).
- **Commitment**: following the proposal of Diab et al. (2009), in this category we have two labels: Commitment and Non-Commitment to indicate the author's stance. Thus, Commitment is used for situations related to present, past, and future situations presented as certain whereas present, past, and future situations, depicted as uncertain, are annotated as Non-Commitment. Only those situations presented with commitment are candidates to be facts.
- **Polarity**: we annotate the whole sentence with respect to polarity. Two tags are used for this level: Positive and Negative. Negative polarity can be used to signal counterfacts.
- **Time**: we differentiate between time and tense. Only time is annotated, and we use three tags: Past, Present and Future. Future events will never be facts even though they might be presented with commitment.

2. Inter-Annotator Agreement

The agreement score and kappa value were assessed over a corpus of 3 newspaper articles, specifically chosen to cover all the possibilities in the tagset. As said above, the number of predicates eligible for annotation is the number identified as such by Freeling. Nevertheless, some of them, especially eventive nouns, are errors as will be explained below. The articles annotated for this experiment were:

1. Así alimenta Youtube las teorías que afirman que la Tierra es plana. 'In this way Youtube feeds theories that claim the Earth is flat' - 113 predicates.[3]
2. ¿Qué hacer si se tiene una hipoteca con cláusula suelo? 'What to do if you have a mortgage with a "floor clause"? - 143 predicates.[4]
3. Los Juegos Olímpicos de la Juventud de 2022 se celebrarán en Dakar. 'The 2022 Youth Olympic Games will be held in Dakar'- 24 predicates.[5]

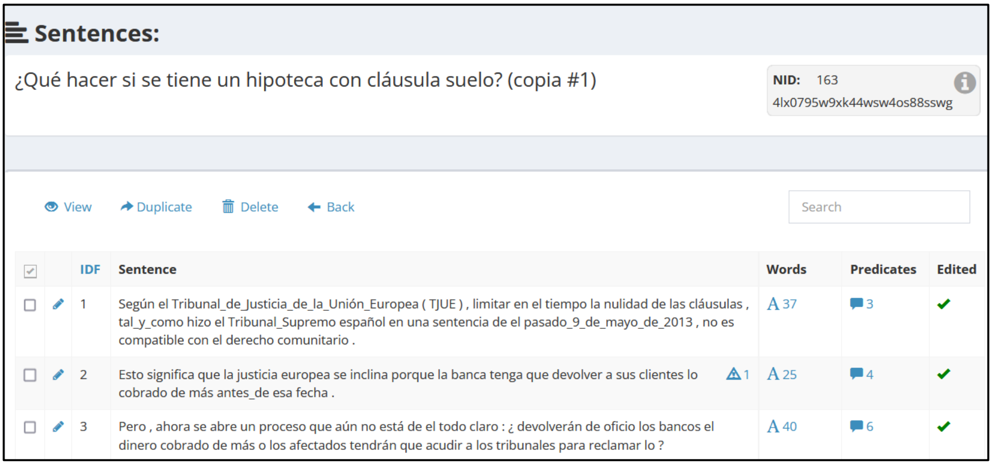
The initial number of predicates was 280 per person, which adds up to a total of 1,680 predicates to be annotated. The criteria used for the annotation was written and discussed a priori. Then, a test was carried out by all 6 annotators in order to ensure the criteria had been clearly understood. After this test, the criteria were discussed again and modifications made.

For this phase of the project, it was decided to deal only with verbs, so eventive nouns were not tagged. Nevertheless, the annotators marked those names of a truly eventive nature to differentiate them from errors inherited from the parsing phase. All errors due to a misanalysis of Freeling were discarded.

2.1. Annotation procedure

The process of annotation took approximately 6 hours per person; the time was tracked by each annotator independently. As a general criterion, annotators had to base their decisions on linguistic clues trying not to use world knowledge, although this is not always possible.

A tool was created to serve a double purpose: firstly, the creation of the corpus and, secondly, the annotation of the predicates with the 5 categories presented above (Fernández-Montraveta, Curell, Vázquez & Castellón, 2020).



Sentences:

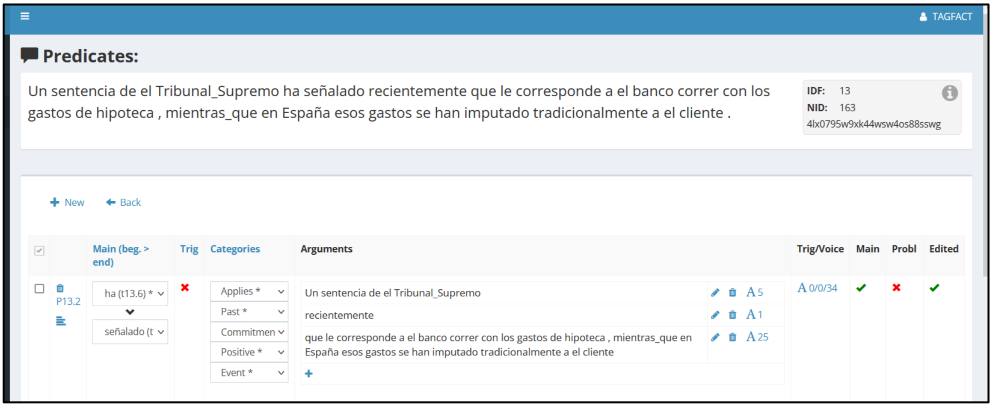
¿Qué hacer si se tiene un hipoteca con cláusula suelo? (copia #1)

NID: 163
4lx0795w9xk44wsw4os88sswg

☑		IDF	Sentence	Words	Predicates	Edited
□		1	Según el Tribunal_de_Justicia_de_la_Unión_Europea (TJUE) , limitar en el tiempo la nulidad de las cláusulas , tal_y_como hizo el Tribunal_Supremo español en una sentencia de el pasado_9_de_mayo_de_2013 , no es compatible con el derecho comunitario .	A 37	3	✓
□		2	Esto significa que la justicia europea se inclina porque la banca tenga que devolver a sus clientes lo cobrado de más antes_de esa fecha . ⚠ 1	A 25	4	✓
□		3	Pero , ahora se abre un proceso que aún no está de el todo claro : ¿ devolverán de oficio los bancos el dinero cobrado de más o los afectados tendrán que acudir a los tribunales para reclamar lo ?	A 40	6	✓

Figure 1. Segmentation of sentences with the associated information.

Figure 1 shows the segmentation of sentences as presented to the annotators. As can be seen, for each sentence, the screen shows the state (edited or unedited), the number of words, and the number of predicates. Once a sentence is selected, the annotator accesses all its predicates and proceeds with the annotation process. The possibilities of selecting trigger words and changing the main verb are also presented to the annotators (Figure 2). a sentence has been finished, the state is changed to Completed.



Predicates:

Un sentencia de el Tribunal_Supremo ha señalado recientemente que le corresponde a el banco correr con los gastos de hipoteca , mientras_que en España esos gastos se han imputado tradicionalmente a el cliente .

IDF: 13
NID: 163
4lx0795w9xk44wsw4os88sswg

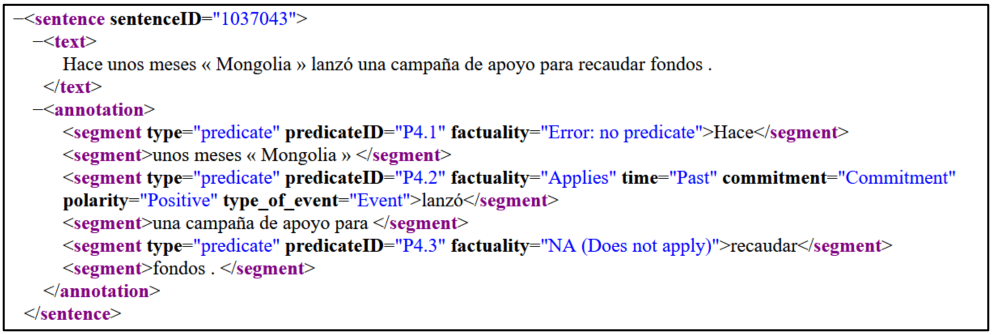
+ New ← Back

☑		Main (beg. > end)	Trig	Categories	Arguments		Trig/Voice	Main	Probl	Edited
□	P13.2	ha (t13.6) * señalado (t	✗	Applies * Past * Commitmen Positive * Event *	Un sentencia de el Tribunal_Supremo recientemente que le corresponde a el banco correr con los gastos de hipoteca , mientras_que en España esos gastos se han imputado tradicionalmente a el cliente +	A 5 A 1 A 25	A 0/0/34	✓	✗	✓

Figure 2. Tool for annotation.

As can be seen in Figure 2, the contraction *del* has been split into 2: *de el*. This is so because Freeling splits all the contractions when lemmatizing the text (preposition + article). This can also be seen in the contraction *al*, which is rewritten as the preposition *a* and the article *el*.

Finally, Figure 3 shows the final annotation result in XML format, as can be downloaded. In the figure, each predicate has its attributes marked in bold and the associated factuality values in blue.



```
–<sentence sentenceID="1037043">
  –<text>
      Hace unos meses « Mongolia » lanzó una campaña de apoyo para recaudar fondos .
    </text>
  –<annotation>
      <segment type="predicate" predicateID="P4.1" factuality="Error: no predicate">Hace</segment>
      <segment>unos meses « Mongolia » </segment>
      <segment type="predicate" predicateID="P4.2" factuality="Applies" time="Past" commitment="Commitment"
      polarity="Positive" type_of_event="Event">lanzó</segment>
      <segment>una campaña de apoyo para </segment>
      <segment type="predicate" predicateID="P4.3" factuality="NA (Does not apply)">recaudar</segment>
      <segment>fondos . </segment>
    </annotation>
 </sentence>
```

Figure 3. Sample of the final annotation in XML format.

2.2. *Assessment of Inter-Annotator Agreement*

As said above, a total of 1,680 annotations were considered (280 predicates per annotator). Only 724 predicates were labeled with the tag Applies, which means that the rest, 956 predicates, were not annotated for a series of reasons: 281 described wishes or desires and were, therefore, not relevant for the description of factuality, 150 were eventive nouns (discarded for reasons explained above) and, finally, 525 items were errors caused by an incorrect tagging (Freeling). These errors were caused by the incorrect detection of the predicate; for example, in the case of past participles, or because grammar is interpreted differently, e.g., periphrastic verbs are annotated in

Freeling as independent predicates but are considered part of the main verb in our project.

A summary of tags used in the 724 predicates is presented below (Table 1). As observed, there seems to be a consistent pattern in the labels used, which could clearly be related to the nature of the journalistic register. Newspaper articles usually narrate events that have happened and, therefore, most of the situations are events presented with commitment that usually express an affirmation over an event that happened in the past (occasionally, it could be an event that has not happened yet).

Table 1. Total number of annotations by category.

Polarity	Positive	Negative			
	690	34			
Time	Past	Present	Future		
	411	276	37		
Commitment	Commitment	Qual.-commitment	Non- commit.	None	
	699	6	18	1	
Eventual types	Event	Property-non-event	Property-Event	Mental	Prop-Abs
	586	84	29	11	14

The agreement score was calculated from the data presented in Table 1. In all of the cases, it was calculated for each pair (6 annotators, a total of 15 pairs). Overall, considering all categories, the agreement score obtained is 88% (see Table 2 below for a detailed account).

Table 2. Average score and range per category.

	Average	**Min.**	**Max.**
Apply	85,57 %	70,55%	93,1%
Time	81,28%	66,04%	94,2%
Polarity	97,27%	93,18%	98,8%
Commitment	99,55%	97,7%	100%
Eventual types	76,45%	69,23%	84,78%

Table 2 presents the minimum and maximum scores for each category. As can be seen, and generally speaking, agreement is rather high. The categories Apply and Time show a lower score when compared with Polarity and Commitment, while the category Event types presents the lowest agreement. The high number of events labeled as Commitment can be easily explained by the fact that we are annotating news updates and, therefore, this is the stance most commonly found. This would not necessarily be the case of other sections of a newspaper such as op-eds. In the case of Polarity, there are just two categories, positive and negative, that are formally marked, which makes agreement more plausible than in other categories.

Figure 4 shows these results graphically. As can be observed, some pairs show higher agreement than others and the category with the highest percentage is

Commitment. The three categories where agreement is more dispersed are Applies, Time and Event Type. Nevertheless, in the case of the former 2 (Applies and Time), agreement varies greatly depending on the pairs compared. Thus, pairs 20 and 26 or 23 and 26 seem to interpret time in a similar way (93,18% and 89,13% respectively) whereas in pairs 20 and 27 a different interpretation of time is observable (66,04%). In this category, many of the differences were due to a different interpretation of indirect speech expressed in the present tense. The lowest agreement rate for all annotators is found in the interpretation of the Event Type, which could be the result of the complexity and the existence of a higher number of tags (*i.e.*, event, state, absolute truth, eventive state).

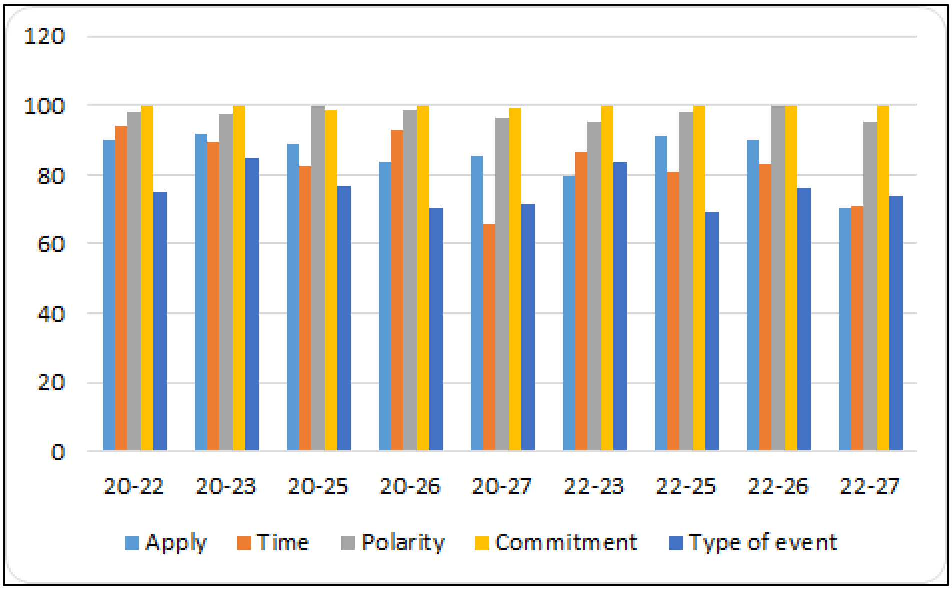


Figure 4. Agreement for all the categories and all the pairs of annotators.

Finally, the choice between the labels Applies and NA- Does not apply has also been proven problematic. In general, the differences observed in the use of these two labels are due to a different interpretation of what can be considered a wish or desire, and also to the interpretation of modality as describing the real world.

As mentioned above, in order to measure inter-annotator reliability, we have used Cohen's kappa (Carletta, 1996). In general, Cohen's kappa (a) is perceived as a more robust measure than descriptive statistics since ϰ also measures by-chance agreement.

$$k=\frac{p_o-p_e}{1-p_e}$$

This measure relates the relative agreement observed between judges (po) and the hypothetical probability according to chance (pe), that is obtained by calculating the probabilities that each observer annotates each category at random from the observed data.

In our case, the ϰ value was calculated in pairs and, then, the average for all the ϰ values calculated. An online calculator[6] was used. Table 3 presents the average values obtained as well as the range (maximum and minimum values) within each category. In the case of the category commitment, the calculation of the average was not possible since some of the pairs presented complete agreement.

Table 3. Average κ values including the maximum and minimum values within each category.

	Apply	Time	Polarity	Commitment	Event Type
Average	0,61	0,64	0,55	nc*	0,35
Max. value	0,8	0,89	1	1	0,57
Min. value	0,2	0,37	0	0,66	0,06

As can be observed, the categories Apply and Time show a slightly over moderate kappa (.41 to .60),[7] whereas the category Polarity presents a slightly lower moderate kappa whereas Event Type presents a fair kappa (.21 - .40). This difference can also be appreciated in the variations observed in the maximum values. Regarding the lowest values, they show that Apply and Event Type are the most complex categories (Figure 5). The problem is different in each of them. The former probably requires further specification whereas the latter is the one that presents the highest number of subcategories.

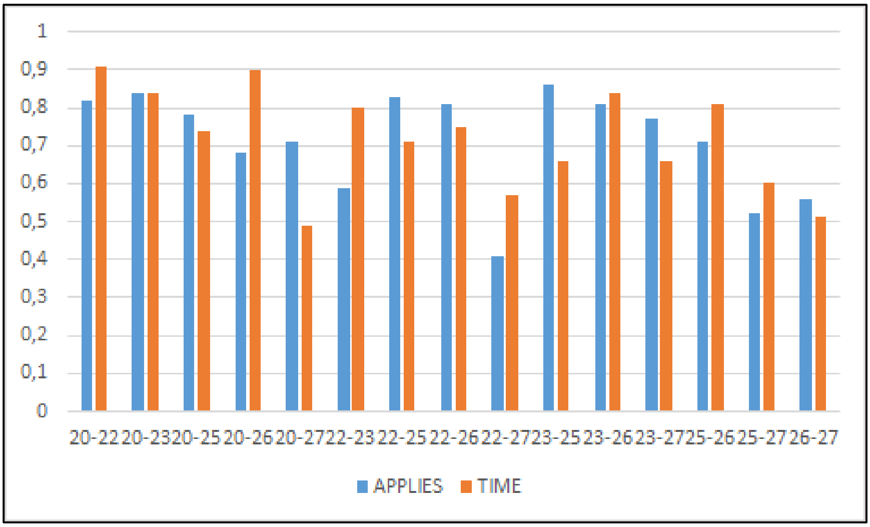


Figure 5. κ values for the category Apply and Event Type for each pair of Annotators.

The characterization of predicates from the point of their eventive structure is the most complex task. This is probably due to two factors: firstly, this category has 5 labels (more than any other and 3 of them are used to specify a different type of state). This difference is not always clear-cut. In fact, in the meetings held to set the criteria, it was always the most problematic category. Disagreement is caused by the fact that not all the annotators interpret sentences in the same way; especially problematic is the interpretation of some periphrastic verbs such as *tratan de mostrar* (try to show) or *podría estar desempeñando* (could be performing).

As already mentioned, agreement on Time shows a moderate kappa while Polarity a weak kappa. In the case of Polarity, the explanation is slightly different, negative values are rather infrequent and if an annotator oversees just one or two cases, the kappa value falls drastically. In sentences as (3), the fact that the adverb *no* is not marked as negative can only be explained as a lapse:

(3) Según esta experta, el hecho de creer que la Tierra es plana "no es necesariamente dañino".
'According to this expert, believing that the Earth is flat "is not necessarily harmful".'

Regarding Time, the situation is slightly different. Many of the differences were caused because some criteria were not explicit enough. For example, some annotators considered time and tense as the same in verbs of communication. When used in the present tense, in written news, usually they refer to a past event of communication. This is the case of the verb *alerta* (warns) in (4):

(4) Landrum alerta que el algoritmo que sugiere nuevos vídeos a las personas que buscan información sobre este tema….
'Landrum warns that the algorithm that suggests new videos to people searching for information on this topic…'

Finally, the category Commitment poses a different kind of problem. In this category, 3 possibilities were predicted (see Section 2.2) one of which, Qualified Commitment, was established for those situations that presented an emphasizer (for example, *seguramente* -in all probability, surely) of the commitment. It was understood that such an emphasizer was acting as a modulator. The problem is that no annotator has used this value and, therefore, there are no values for this category.

As a sample of the differences in annotation, Table 4 presents the confusion matrices of a pair of evaluators for the five categories annotated. In general, judges obtained similar levels of agreement that varied depending on the category. However, among all the peers there is a judge who always obtains the worst index of agreement, judge 27, as seen in Figure 4.

Table 4. Confusion matrix of one pair of judges for all four categories.

Time	Past	Present	Future
Past	44	1	0
Present	5	35	0
Future	0	0	3

Polarity	Positive	Negative
Positive	84	1
Negative	0	3

Commitment	Commitment	Qual.Comm.	Non-Comm
Commitment	87	0	0
Qual. Comm.	0	1	0
Non-Comm.	0	0	0

Eventual Types	Event	Prop.non-event	Prop.event	Mental	Prop-Abs
Event	56	1	1	1	2
Prop.non-event	0	0	0	0	0
Prop.event	5	2	0	0	0
Mental	0	0	0	0	4
Prop-Abs	10	0	0	0	6

All things considered, and after analyzing the differences, it was concluded that some of them are due to attention mistakes (lapses). This type of error is easy to understand since in our scheme, there are a total of 15 choices, distributed into 5 categories, and the higher the complexity of the annotation task the more likely it is that this kind of mistake will be made. Another general error comes from the a priori decision of assigning default values, a decision taken with the idea of simplifying the annotation, but that proved to be wrong in the end since it has been a source of errors.

From the data presented and analyzed in this article, it was concluded that the annotation strategy applied was impractical. In other words, it was not optimal since too many attention errors were detected. For this reason, we decided to change the procedure for the annotation of the GS so that each annotator would specialize in just one category. This decision was taken, not for a lack of confidence in the expertise of the annotators, but because annotating 5 categories (15 choices) has proved to be complicated for the annotators.

CONCLUSIONS

This paper has presented the inter-annotator agreement of the TAGFACT Agreement Corpus. The TAGFACT project aims to annotate factuality as presented by the narrator using a multi-level annotation scheme. This study was performed in order to help see how well categories were delineated as well as how trustworthy and replicable the annotation is.

The data analyzed in this paper shows the annotation of 3 pieces of news by six annotators. It has been assessed using Cohen's kappa and descriptive statistics. After reviewing the agreement figures, we concluded that even though the numbers are rather good, it is not enough but we need to take into account that the annotation task of semantic categories is not an easy task. As seen in Section 3.2, some further specifications are needed in the description of the general criteria. The first decision Apply versus NA- Does not apply presents the first major problem. Sometimes the line between real (or probable situations), interpretations, and unreal situations is blurred. Moreover, since we annotate the degree of certainty, probable or future situations can also be presented with commitment. Therefore, this concept needs to be further specified.

The second category that needs major revision is Event Type. In order to solve this problem, we have opted for specialization in the annotation process, so it is ensured that the criteria are coherently applied. Finally, the last category that needs revision is Commitment. In this case, 3 labels were first considered but it was later decided to keep just 2 Commitment and Non-commitment since Qualified Commitment proved to be unnecessary and problematic.

Lastly, some improvements need to be done in the software used for annotation so that the system is more useful in the annotation process. The most important change will help prevent further annotation errors due to lapses by adding a controller to let the annotator know when the 'by-default categories' have not been changed. Our future work is to redefine the annotation scheme, using the information provided by the results presented in this paper, finish with the annotation of the GS and the implementation of the program to proceed to the automatic annotation.

The GS, the final corpus and the annotator will be made freely available to the community through the research group website and other pertinent institutions (http://grial.edu.es/web/en/downloads-access/).

REFERENCES

Bian, J., Liu, Y., Agichtein, E. & Zha, H. (2008). Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media. In *The International World Wide Web Conference Committee-IW3C2* (pp. 467-476). Beijing: China.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, *22*(2), 249-254.

Diab, M. T., Levin, L., Mitamura, T., Rambow, O., Prabhakaran, V. & Guo, W. (2009). Committed belief Annotation and Tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, *Association for Computational Linguistics* (pp. 68-73). Suntec: Singapore.

Fernández-Montraveta, A. H., Curell, G. Vázquez & I. Castellón (2020). The TAGFACT Annotator and Editor: A Versatile Tool. *Research in Corpus Linguistics*, *8*(1), 131-146.

Hasanain, M., Suwaileh, R., Elsayed, T., Barrón-Cedeño, A. & Nakov, P. (2019). Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality. 4.0). In *CEUR Workshop Proceedings, CEUR-WS, 2019, 2380* (pp. 1-15). Lugano, Switzerland.

Landis, J. R. & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1),159-174.

Leblay, J. A. (2017). A Declarative Approach to Data-Driven Fact Checking. In *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 147-153). San Francisco, California: US.

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). Lisbon, Portugal.

Lloberes M., Castellón, I. & Padró, L. (2015). Suitability of ParTes Test Suite for Parsing Evaluation. *Proceedings of the 14th International Conference on Parsing Technologies.* Association for Computational Linguistics (pp. 61-65). Bilbao, España.

Matsuyoshi, S., Eguchi, M., Sao, Ch., Murakami, K., Inui, K. & Matsumoto, Y. (2010). Annotating Event Mentions in Text with Modality, Focus, and Source Information. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation-LREC'10* (pp. 1456-1463). Valletta, Malta.

Marneffe, M. C., Manning, C. D. & Potts, C. (2012). Did it Happen? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics*, *38*(2), 301-333.

Minard, E., Speranza, E. & Caselli, T. (2016). The EVALITA 2016 Event Factuality Annotation Task (FactA). In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016)* & *Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop-EVALITA 2016* (pp. 36-39). Napoli, Italy.

Narita, K., Mizuno, J. & Inui, K. (2013). A Lexicon-based Investigation of Research Issues in Japanese Factuality Analysis. In *International Joint Conference on Natural Language Processing* (pp. 587-595). Nagoya, Japan.

Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *International Conference on Language Resources and Evaluation-LREC2012* (pp. 2473-2479). Istanbul, Turkey.

Sahu, I. & Majumdar, D. (2017). Detecting Factual and Non-Factual Content in News Articles. In CODS '17: *Proceedings of the Fourth ACM IKDD Conferences on Data Sciences* (pp. 1-12), Chennai, India.

Santana, A., Nieuwenhuijsen, D., Spooren, W. & Sanders, T. (2017). Causality and Subjectivity in Spanish Connectives: Exploring the Use of Automatic Subjectivity Various Text Types. Discours, Revue de linguistique, psycholinguistique et informatique, 20 [online]. Retrieved from: http://journals.openedition.org/discours/9307

Saurí, R. (2008). *A Factuality Profiler for Eventualities in Tex*t. Ph.D. Thesis, Brandeis University, United States.

Saurí, R. & Pustejovsky, J. (2009). FactBank: A Corpus Annotated with event Factuality. *Language Resources and Evaluation*, *43*(3), 227-268.

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.

Soni, S., Mitra, T., Gilbert, E. & Eisenstein, J. (2014). Modeling Factuality Judgments in Social Media Text. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 415-420). Baltimore, Maryland.

Tianxiong, H., Peifeng, L. & Qiaoming, Z. (2018). Identifying Chinese Event Factuality with Convolutional Neural Networks. In Y. Wu, J. F. Hong & Q. Su (Eds.), *Chinese Lexical Semantics. CLSW 2017*. Lecture Notes in Computer Science 10709 (pp. 284-292). Springer: Cham.

van Son, C., M. van Erp, Fokkens, A. & Vossen, P. (2014). Hope and Fear: Interpreting by Integrating Sentiment and Event Factuality. In *Proceedings of the 9th Language Resources and Evaluation Conference-LREC2014* (pp. 3587-3864). Reykjavik, Iceland.

Vázquez, G., A. & Fernández-Montraveta, A. H. (2020). Annotating Factuality in the Tagfact Corpus. M. Fuster-Márquez, C. Gregori-Signes & J. Santaemilia Ruiz (Eds.), *Multiperspectives in Analysis and Corpus Design* (pp. 115-125). Granada: Comares.

Vendler, Z. (1957). Verbs and Times, *The Philosophical Review*, *66*(2),143-160.

Vigus, M., Van Gysel, J. E. L. & W. Croft. (2019). A Dependency Structure Annotation for Modality. *Proceedings of the First International Workshop on Designing Meaning Representations* (pp. 182-198). Florence, Italy.

Wiebe, J. & Riloff, E. (2011). Finding Mutual Benefit between Subjectivity Analysis and Information Extraction. *IEEE Transactions on Affective Computing*, *2*(4),175-191.

Wonsever, D., Malcuori, M. & Rosá Furman, A. (2008). Sibila: Esquema de anotación de eventos. Reportes Técnicos [en línea]. Disponible en: https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/3419

Wonsever, D., Rosá, A., & Malcuori, M. (2016). Factuality Annotation and Learning in Spanish Texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation-LREC 16* (pp. 2076-2080). Portorož, Slovenia.

NOTES

[1] Beatriz Escudero (PP) a Gabriel Rufián: "No me guiñes el ojo, imbécil" (elperiodico.com).

[2] Supremo de EEUU respalda a pastelero que no quiso hacer tarta para pareja gay (elperiodico.com).

[3] La Vanguardia 19/02/2019.

[4] El Periódico 21/12/2016.

[5] La Vanguardia 8/10/2018.

[6] http://vassarstats.net/kappa.html

[7] In order to decide the thresholds for the kappa value we use the interpretation provided by Landis and Koch (1977).

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of this work (FFI2017-84008-P). We would also like to thank the six annotators for their help with this experiment.