ANNOTATION GUIDE

GEMECO PROJECT

Visibilization of gender gaps in media

SINAI "SISTEMAS INTELIGENTES DE ACCESO A LA INFORMACIÓN"

TABLE OF CONTENTS

1. INTRODUCTION

This document describes the annotation guide used to generate the annotated corpus in Spanish that will be used to develop the GeMeCo system: Visibilization of Gender Gaps in the Media, based on the *GenderGapTracker*[1].

It is a real-time automated system that measures the ratio of men to women quoted in major Canadian media outlets. The Spanish version will use articles published in six Spanish media: El Mundo, ABC, El País, La Vanguardia, Público and El Periódico. This guide is based on the *GenderGapTracker* annotation guide used for English, adapting it to Spanish.

2. ANNOTATION GUIDE

2. 1. PRELIMINARY ISSUES

The core of the annotation is always the quotation (**QUOTATION**) which must necessarily have an entity (**REFERENCE**) to which the quotation can be associated. Therefore, whenever we annotate we must always have two elements: **REFERENCE** and **QUOTATION**. That is a quotation and someone to whom to attribute it.

The elements to be annotated are **QUOTATION** (the quotation itself, either direct (textual) or indirect (paraphrased)) and its corresponding **REFERENCE** (the entity to which the quotation is attributed). In this way, only and exclusively quotations that can be attributed to some entity present in the text are annotated. For this purpose, we can rely on *dicendi*, declarative, or speech verbs (to say, to argue, to allege, to narrate, to preach,...).[2]

Other elements surrounding the appointment are **SPEAKER**, **VERB,** and **TRIGGER**. All of them are related to **QUOTATION** (all links or relationships established between the quotation and other elements start from the quotation). All these elements mentioned should be annotated if any.

It is also important to differentiate whether they are words spoken by the entity to which the quote belongs or whether they are the journalist's own words.

In addition, the maximum length possible in a quotation will be noted. Example: *Llegó a confesar que* **[*la ruptura le había hecho sufrir más dolor que el fallecimiento de Carlos Falcó por la incertidumbre de la situación*]**.[3]

[1] https://gendergaptracker.informedopinions.org/

[2] Consult the ADESSE database.

[3] She confessed that the breakup had made her suffer more pain than the death of Carlos Falcó because of the uncertainty of the situation.

On the other hand, it should be noted that the period is not noted in the quote. In the following example, the full stop is left out of the quotation: *la viuda de Carlos Falcó defendía que* **[*no era conocedora de la ruptura ya que solo recibió un WhatsApp de dos líneas al que no dio mayor importancia*]**.[4]

2.1.1. Data format

Each quotation should have a record as illustrated below. We will use this sentence as an example: *Mientras que el juez de la Audiencia Nacional aseguraba que habían roto el 12 de agosto*.[5]

Annotation:

```
{
“speaker”: “el juez de la Audiencia Nacional”
“verb”: “aseguraba”
“disparador”: “”
“quote”: “habían roto el 12 de agosto”
“speaker_index”: “(x,x+32)”
“verb_index”: ”(y, y+9)”
“quote_index”: “(z,z+27)”
“reference”: “Santiago Pedraz”
“gender”: “m”
}
```

The "gender" field has three possible values {m,f,u}.

2.2. REFERENCE

The first time the entity appears in the text is noted in full, i.e., with first and last names, to which at least one citation can be attributed. Its gender must be identified: **MALE**, **FEMALE** or **UNKNOWN**. The first two will be assigned to male or female names that appear as entities with attributable citations. The **UNKNOWN** label will be assigned to those entities that do not have a biological gender such as newspapers, names of institutions, political parties, etc.

[4] Carlos Falcó's widow defended that she was not aware of the breakup since she only received a two-line WhatsApp to which she did not attach much importance.

[5] While the judge of the Audiencia Nacional assured that they had broken up on August 12.

In addition, if an entity (**REFERENCE**) appears several times in the text, only the first time it appears is annotated, therefore, the citations attributed to it will be linked only to this first annotated appearance of the **REFERENCE**.

There may be cases where the first occurrence of the **REFERENCE** is given as a complement to the name in the apparent form of **SPEAKER**. In this case, since the **REFERENCE** appears after the supposed **SPEAKER**, the latter would not be annotated:

- *"Por lo tanto, siempre trabajamos en la excelencia"*(**QUOTATION**), *destaca*(**VERB**) *la directora general* (**SPEAKER**), *la francesa Anne-Laure Suvage Cayrol*(**REFERENCE**).[6] **REFERENCE** is a complement to the name, as in this example (the name Anne-Laure appears for the first time in the text).

2.2.1. Reference unknown

Citations attributed to companies, groups of people, etc. are noted only in the case of being produced by a specific and identifiable group:

- F*ue entonces, cuando Société Générale* (**REFERENCE-UNKNOWN**), b*anco propietario de la firma, anunció* (**VERB**) s*u intención de adquirir LeasePlan por 4.900 millones de euros* (**QUOTATION**).[7]
- *Desde la Asociación Española de Proveedores de Automoción (Sernauto) aseguran que el 60% de empresas de componentes continíuan afectadas por la falta de semiconductores.*[8] Here the quote is attributed to some unknown persons belonging to the company, that is why we do not write it down.
- "*Durante los primeros meses del año ha habido un retroceso de matriculaciones[...]", constata la directiva.*[9] A document is cited and not a person, we leave it unnoted.
- *PSOE* (**REF.- UNK.**) y *Podemos* (**REF.- UNK.**) *han pedido esperar al Tribunal Supremo y a la Fiscalía para que saquen conclusiones para dar respuesta uniforme a la revisión de las sentencias.*[10]

[6] "Therefore, we always work on excellence", stresses the general manager, French Anne-Laure Suvage Cayrol.

[7] It was then that Société Général, the bank that owns the firm, announced its intention to acquire LeasePlan for 4.9 billion euros.

[8] The Spanish Association of Automotive Suppliers (Sernauto) states that 60% of component companies continue to be affected by the lack of semiconductors.

[9] "During the first months of the year, there has been a decline in registrations," the board notes.

[10] PSOE and Podemos have asked to wait for the Supreme Court and the Prosecutor's Office to draw conclusions in order to give a uniform response to the review of the sentences.

2.3. QUOTATION

It is both a direct and indirect quotation. All the relations to the other elements are derived from it. The quotation starts from the conjunction "que", which will never be noted as part of **QUOTATION**.

- *Mientras que el juez de la Audiencia Nacional aseguraba que <u>habían roto el 12 de agosto</u>* (**QUOTATION**).[11]

2.3.1. Direct quotes

Direct quotations are usually marked with quotation marks (both English and Spanish, depending on the newspaper). They are the easiest to identify and are included in the annotation, as in this example:

- *Llegó a confesar que la ruptura le había hecho sufrir más dolor que el fallecimiento de Carlos Falcó por la incertidumbre de la situación y que <u>"Aún no sé el motivo por el que tomó la decisión. No me ha dado explicaciones, solo la del WhatsApp. Ha sido muy traumático no poder hablar ni que se me expliquen las cosas"</u>* (**QUOTATION**).[12]

However, there are cases where the verb is inserted between the quotation: "---" vb "---". It does not mean that there are two different quotations, because it is the same, only that the journalist, in his writing, has decided to cut it by putting the verb in the middle of it. In this case, everything will be taken as the same quotation, and, in addition, the verb will be noted inside the quotation, which will continue to establish a relationship between the quotation and the verb.

- *"Sin duda alguna", dice, "la generalización de la telemática aplicada a la gestión de flotas supone un paso importante".*[13]

In this example we annotate "Sin duda alguna", dice, "la generalización de la telemática aplicada a la gestión de flotas supone un paso importante" as **QUOTATION**, as well as "dice" as **VERB** overlapping with the previous annotation.

[11] While the judge of the Audiencia Nacional assured that they had broken up on August 12.

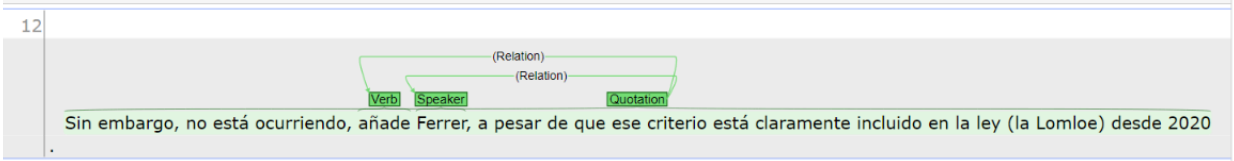
[12] He confessed that the breakup had made him suffer more pain than the death of Carlos Falcó because of the uncertainty of the situation and that "I still don't know why he made the decision. He has not given me explanations, only the WhatsApp one. It has been very traumatic not to be able to talk or to have things explained to me".

[13] "Undoubtedly," he says, "the generalization of telematics applied to fleet management is an important step".

2.3.2. Indirect quotes

We speak of indirect quotation when the words are not reproduced verbatim, which is why they do not appear formally in quotation marks. Example: *Raquel Sánchez destacó este lunes que* [***la incidencia de la huelga convocada por la Plataforma Nacional en Defensa del Transporte por Carretera es "mínima" y reveló que se han desplegado 50.000 agentes para evitar disturbios***] .[14]

In case there is a verb embedded in the quotation, the quotation is annotated including the verb, which is also annotated with the corresponding **VERB** label, within the quotation. In addition, the relationship between **QUOTATION** and **VERB** is still established:



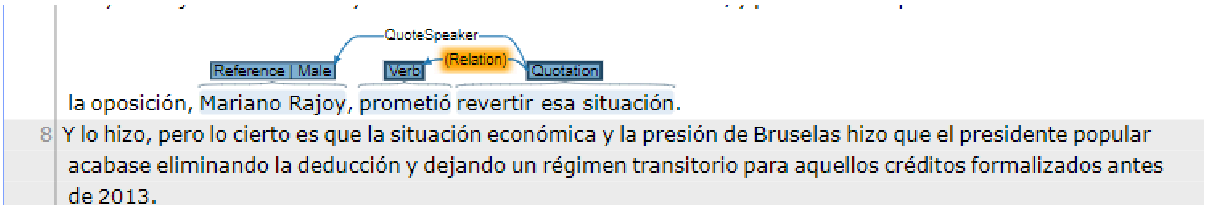
Sin embargo, no está ocurriendo, añade Ferrer, a pesar de que ese criterio está claramente incluido en la ley (la Lomloe) desde 2020 .

Fig. 1: example extracted from manual annotation with the Inception platform.

2.3.2.1. Indirect quotations in the form of complex or subordinate clauses

When we have an indirect quotation made as a subordinate clause dependent on the main verb (**VERB**), the subordinating conjunctive nexus (que) will be taken out of the quotation. For example, *es más, confiesa que* ***le pidió expresamente que también fuera al suyo****, algo que en su momento ocultó.*[15]

In the case of a subordinate clause with an infinitive clause, the quotation will be indicated from the infinitive verb. For example: *Mariano Rajoy* [**REFERENCE**, **MASCULINE**] *prometió* [**VERB**] *revertir esa situación* [**QUOTATION**].[16]



la oposición, Mariano Rajoy, prometió revertir esa situación.

Y lo hizo, pero lo cierto es que la situación económica y la presión de Bruselas hizo que el presidente popular acabase eliminando la deducción y dejando un régimen transitorio para aquellos créditos formalizados antes de 2013.

Fig. 2: example extracted from manual annotation with the Inception platform.

[14] Raquel Sánchez highlighted this Monday that the incidence of the strike called by the National Platform in Defense of Road Transport is "minimal" and revealed that 50,000 agents have been deployed to avoid disturbances.

[15] He even confesses that he expressly asked her to go to his, something he hid at the time.

[16] Mariano Rajoy promised to reverse this situation.

2.3.2.2. Indirect quotations in the form of compound or coordinated sentences

When annotating the text, if we have a quotation syntactically realized as a compound sentence or coordinated with the conjunction "y"[17], we have to point out two different **QUOTATIONS** that will depend on different verbs, although they may share the **SPEAKER**.

For example, *el líder de la oposición comentó* (**VERB**^{1}) *que* ***él no habría seguido esa vía*** *y aseguró* (**VERB**^{2}) *que* ***no estaba de acuerdo con las decisiones tomadas en el congreso***.[18]

Likewise, we must pay special attention because it is very common to find in quotations made as a subordinate clause, within this one, that a coordinated clause is inserted. For example: *Pancho comentó a esta periodista que* ***no sentía que Sabina le hubiese hecho "un feo", y que entendía que no pudiese haber ido a los dos conciertos***.[19]

In this case, the main verb is "comentó", therefore, the whole subordinate clause starting with the conjunction que, would be annotated as a quotation.

2.3.2.3. Quotes within quotes

When an indirect quotation contains an exact quotation within it, the entire indirect quotation is noted:

- *La viuda de Falcó* (**SPEAKER**) *ha confesado* (**VERB**) *que* ***no borraría la etapa que ha vivido junto al juez ya que «es una persona que ha estado en mi vida. Hemos vivido cosas muy bonitas. Ha sido un año fantástico. Por el momento, no lo voy a hacer*** (**QUOTATION**)».**[20]**

2.4. SPEAKER

This element is annotated when we are facing the appearance of the entity in the same sentence where the quotation is found once it has already been mentioned previously, that is to say, the **REFERENCE** has already been annotated in previous lines. It can range from parts of the surname (e.g. Doña, Pedraz, Castro, Roca, Lasso Vega,...), to what we call

[17] And.

[18] The leader of the opposition commented that he would not have followed that path and assured that he did not agree with the decisions made in the congress.

[19] Pancho commented to this journalist that he did not feel that Sabina had done him "a disservice", and that he understood that he could not have gone to both concerts.

[20] Falcó's widow has confessed that she would not erase the stage she has lived with the judge since "he is a person who has been in my life. We have lived very nice things. It has been a fantastic year. For the moment, I'm not going to do it."

epithets (the widow, the manager of Northgate, etc.). These refer directly to the **REFERENCE** already noted above. In addition, pronouns can also function as **SPEAKER**.

The subject represented by the **SPEAKER** is noted in full (the whole nominal syntagm including the article) if it is a common noun. If it is a proper noun, the complements of the noun that it may have are not taken into account:

- *Mientras que el juez de la Audiencia Nacional aseguraba que habían roto el 12 de agosto.*[21]
- *Alexandru Constantin Chituta, del Museo Nacional Brukenthal (Rumanía).* Aquí dejamos fuera la aposición "*del Museo Nacional Brukenthal (Rumanía)*". [22]

In the case of having a **SPEAKER** that refers to two or more of the annotated **REFERENCES** using forms such as both, the couple, the protagonists, etc., they will be annotated as **SPEAKER**. Therefore, as many lines of relationships must be established between the quotation and the entities as there are **REFERENCES** annotated in the **SPEAKER**.

For example: *Esther Doña* (**REFERENCE**) y *Santiago Pedraz* (**REFERENCE)** [...] ***los protagonistas*** **(SPEAKER)** *comunicaron …*[23]

2.4.1. Quotespeaker

Since the tool used does not allow the same text string to be tagged twice, if the entity (**REFERENCE**) is the same as the person (**SPEAKER**), the "QuoteSpeaker" tag will have to be added to the relationship between "**QUOTATION**" and "**REFERENCE**".

This is a tag used when **REFERENCE** and **QUOTATION** are in the same sentence. They are used indistinctly if they are direct quotations or paraphrases.

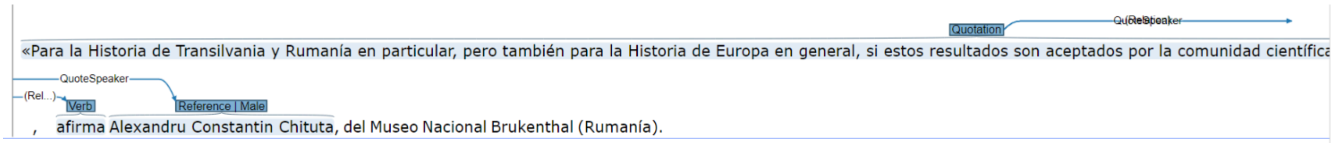
For example: *"Para la Historia de Transilvania y Rumanía en particular, pero también para la Historia de Europa en general, si estos resultados son aceptados por la comunidad científica significará la adición de otra figura histórica importante en nuestra Historia"* (**QUOTATION**)*, afirma* (**VERB**) *Alexandru Constantin Chituta* (**REFERENCE, MALE**)*,*

[21] While the judge of the Audiencia Nacional assured that they had broken up on August 12.

[22] Alexandru Constantin Chituta, from the Brukenthal National Museum (Romania). Here we leave out the apposition "from the Brukenthal National Museum (Romania)".

[23] Esther Doña and Santiago Pedraz [...] the protagonists communicated …

del Museo Nacional del Brukenthal (Rumanía).[24]



«Para la Historia de Transilvania y Rumanía en particular, pero también para la Historia de Europa en general, si estos resultados son aceptados por la comunidad científica

, afirma Alexandru Constantin Chituta, del Museo Nacional Brukenthal (Rumanía).

Fig. 3: example extracted from manual annotation with the Inception platform.

2.5. VERB

*An element that need not always appear. Only the main verb is noted. Examples: He came to **confess**..., ...had been **confessed**...*

2.6. TRIGGER

They act as "announcers" that a quote will be given in the text. For example: ***Según*** [**TRIGGER**] *García Roji* [**SPEAKER**]*, actualmente "un 35% del parque móvil no tiene el etiquetado que se exigirá en estas zonas"* [**QUOTATION**],...[25]

This one works as an "announcer" because, without the need for a declarative verb, we know that a date is coming. Thus, in the case of finding the following: *"nuestra relación es imposible, hablamos algún día, cuídate y besos"* *según* ***contó*** *poco después la viuda del Marqués de Griñón.*[26] We will not annotate *según*[27] as **TRIGGER**, since there is a verb (*contó*) that introduces the quote and, therefore, it is not acting as such.

2.6.1. List of TRIGGER

The following **TRIGGERs** are included in this list:

- De acuerdo con (According to)
- Según (According to)
- Teniendo en cuenta a (Based on)
- Con base en (Based on)
- Citando a (Citing)
- Tal como (As)

[24] "For the History of Transylvania and Romania in particular, but also for the History of Europe in general, if these results are accepted by the scientific community it will mean the addition of another important historical figure in our History", says Alexandru Constantin Chituta, from the Brukenthal National Museum (Romania).

[25] According to García Roji, currently "35% of the vehicle fleet does not have the labeling that will be required in these zones", ...

[26] "Our relationship is impossible, we will talk someday, take care and kisses" as told shortly after the widow of the Marquis of Griñon.

[27] According to.

- Como lo hace notar (As noted)
- Empleando las palabras de (Using the words of)
- A juicio de (In the opinion of)
- Desde la posición de (From the position of)
- Dicho con palabras de (In the words of)
- En la opinión de (In the opinion of)
- Son de opinión (Are of the opinion of)
- Desde el punto de vista de (From the point of view of)
- En su opinión (In their opinion)
- Acorde a (According to)
- En palabras de (In the words of)