CORPUS PRESEEA

PROYECTO PARA EL ESTUDIO SOCIOLINGÜÍSTICO DEL ESPAÑOL DE ESPAÑA Y DE AMÉRICA

Manual de consulta

https://preseea.uah.es

F. Javier Pueyo Mena
Francisco Gago-Jover

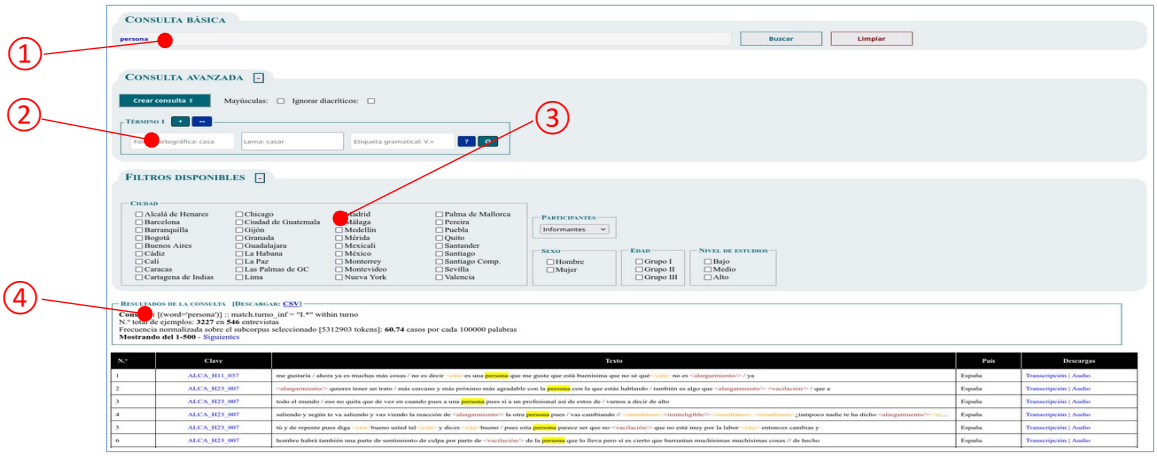
Versión 1.0
Febrero 2025

Tabla de contenidos

A. Interfaz de consulta

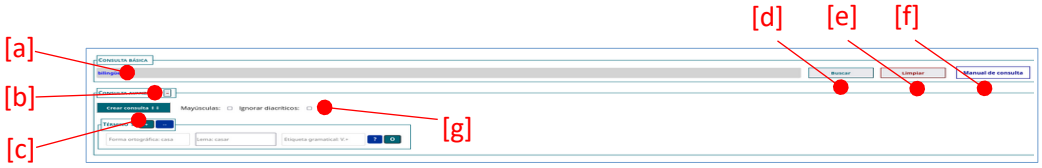
La interfaz de consulta del corpus *Preseea* tiene cuatro componentes principales:

1 Buscador
2 Generador de consultas (Consulta avanzada)
3 Filtros
4 Resultados



1 Buscador (consulta básica) ①

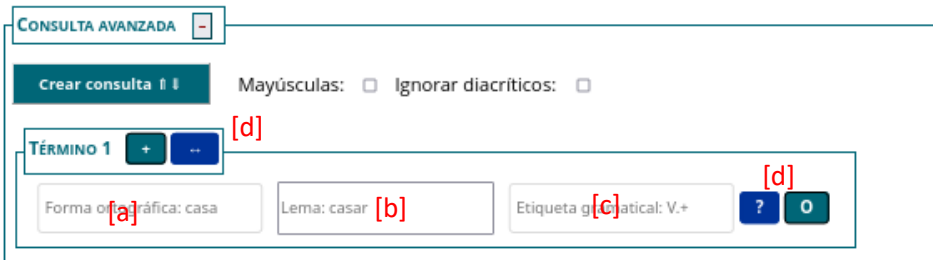
Situado en la parte superior de la interfaz, encontramos los siguientes elementos:



a Casilla de entrada de datos: los términos de la consulta pueden escribirse directamente en esta casilla (ver el ***Apéndice A*** para una explicación detallada de la sintaxis de las búsquedas) o bien pueden ser generados de forma semi-automática con el **generador de consultas** de la consulta avanzada ②.
b Botón para ocultar o mostrar las opciones avanzadas de generación de consultas.
c Botón para crear e insertar la consulta
d Botón para lanzar la consulta
e Botón para limpiar la casilla de entrada de datos
f Botón para descargar el *Manual de consulta*
g Casillas de selección de mayúsculas y diacríticos

2 Generador de consultas (consulta avanzada) ②

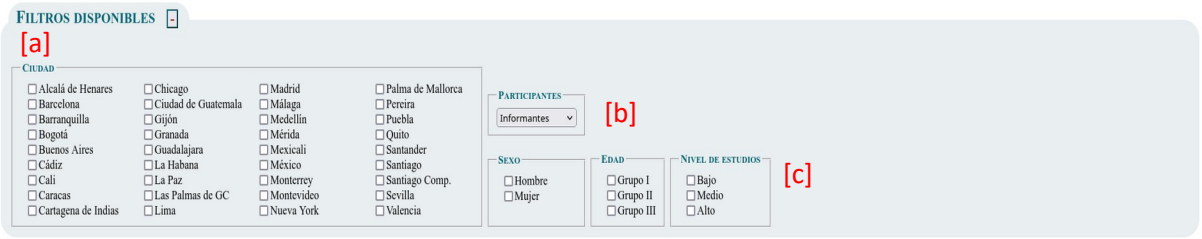
Encontramos los siguientes elementos:



a Forma ortográfica
b Lema
c Etiqueta gramatical
d Botones
- + Añadir término
- ↔ Añadir operador de repetición
- ? Definir etiqueta
- O Añadir alternativa

3 Filtros ③

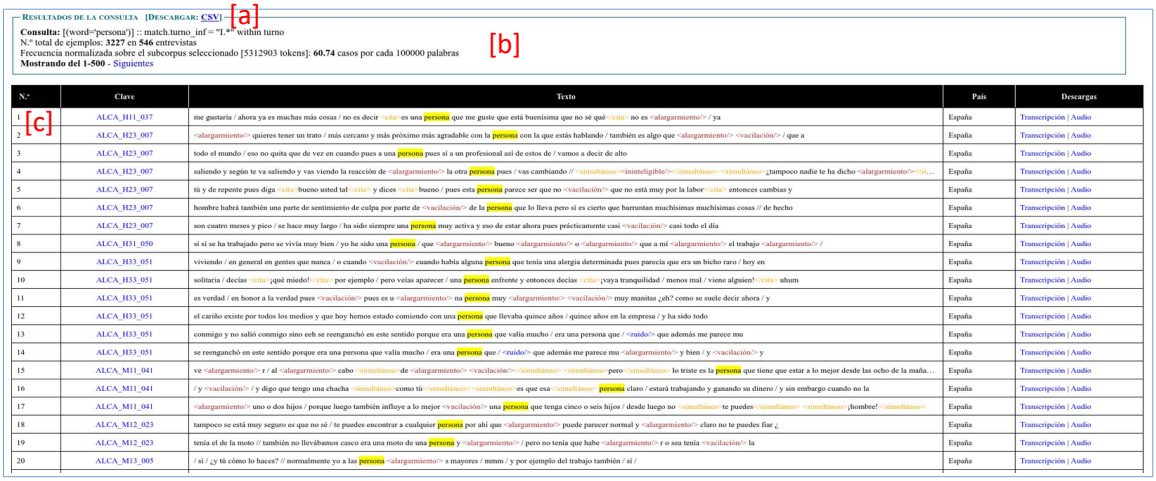
En la sección correspondiente a los filtros encontramos una serie de menús desplegables que permiten el filtrado de los resultados.



- [a] **Ciudad**: casillas de selección múltiples
- [b] **Participantes**: menú desplegable (Informantes, entrevistadores o ambos)
- [c] **Sexo**, **edad** y **nivel de estudios**: casillas de selección múltiples

4 Resultados ④

Una vez efectuada la consulta los resultados se muestran en la parte inferior de la interfaz, donde encontramos los siguientes elementos



RESULTADOS DE LA CONSULTA [DESCARGAR: CSV]

Consulta: [(word='persona')] :: match.turno_inf = "I.*" within turno
N.º total de ejemplos: **3227** en **546** entrevistas
Frecuencia normalizada sobre el subcorpus seleccionado [5312903 tokens]: **60.74** casos por cada 100000 palabras
Mostrando del 1-500 - Siguientes

N.º	Clave	Texto	País	Descargas
1	ALCA_H11_037	me gustaría / ahora ya es muchas más cosas / no es decir <cita>es una persona que me guste que está buenísima que no sé qué</cita> no es <alargarmiento/> / ya	España	Transcripción | Audio
2	ALCA_H23_007	<alargarmiento/> quieres tener un trato / más cercano y más próximo más agradable con la persona con la que estás hablando / también es algo que <alargarmiento/> <vacilación/> / que a	España	Transcripción | Audio
3	ALCA_H23_007	todo el mundo / eso no quita que de vez en cuando pues a una persona pues sí a un profesional así de estos de / vamos a decir de alto	España	Transcripción | Audio
4	ALCA_H23_007	saliendo y según te va saliendo y vas viendo la reacción de <alargarmiento/> la otra persona pues / vas cambiando // <simultáneo><ininteligible/></simultáneo> <simultáneo>¿tampoco nadie te ha dicho <alargarmiento/></si…	España	Transcripción | Audio
5	ALCA_H23_007	tú y de repente pues diga <cita>bueno usted tal</cita> y dices <cita>bueno / pues esta persona parece ser que no <vacilación/> que no está muy por la labor</cita> entonces cambias y	España	Transcripción | Audio
6	ALCA_H23_007	hombre habrá también una parte de sentimiento de culpa por parte de <vacilación/> de la persona que lo lleva pero sí es cierto que barruntan muchísimas muchísimas cosas // de hecho	España	Transcripción | Audio
7	ALCA_H23_007	son cuatro meses y pico / se hace muy largo / ha sido siempre una persona muy activa y eso de estar ahora pues prácticamente casi <vacilación/> casi todo el día	España	Transcripción | Audio
8	ALCA_H31_050	sí sí se ha trabajado pero se vivía muy bien / yo he sido una persona / que <alargarmiento/> bueno <alargarmiento/> o <alargarmiento/> que a mí <alargarmiento/> el trabajo <alargarmiento/> /	España	Transcripción | Audio
9	ALCA_H33_051	viviendo / en general en gentes que nunca / o cuando <vacilación/> cuando había alguna persona que tenía una alergia determinada pues parecía que era un bicho raro / hoy en	España	Transcripción | Audio
10	ALCA_H33_051	solitaria / decías <cita>¿qué miedo!</cita> por ejemplo / pero veías aparecer / una persona enfrente y entonces decías <cita>¡vaya tranquilidad / menos mal / viene alguien!</cita> uhum	España	Transcripción | Audio
11	ALCA_H33_051	es verdad / en honor a la verdad pues <vacilación/> pues es u <alargarmiento/> na persona muy <alargarmiento/> <vacilación/> muy manitas ¿eh? como se suele decir ahora / y	España	Transcripción | Audio
12	ALCA_H33_051	el cariño existe por todos los medios y que hoy hemos estado comiendo con una persona que llevaba quince años / quince años en la empresa / y ha sido todo	España	Transcripción | Audio
13	ALCA_H33_051	conmigo y no salió conmigo sino eeh se reenganchó en este sentido porque era una persona que valía mucho / era una persona que / <ruido/> que además me parece mu	España	Transcripción | Audio
14	ALCA_H33_051	se reenganchó en este sentido porque era una persona que valía mucho / era una persona que / <ruido/> que además me parece mu <alargarmiento/> y bien / y <vacilación/> y	España	Transcripción | Audio
15	ALCA_M11_041	ve <alargarmiento/> r / al <alargarmiento/> cabo <simultáneo> de <alargarmiento/> <vacilación/></simultáneo> <simultáneo> <simultáneo>pero</simultáneo> lo triste es la persona que tiene que estar a lo mejor desde las ocho de la maña…	España	Transcripción | Audio
16	ALCA_M11_041	/ y <vacilación/> / y digo que tengo una chacha <simultáneo> como tú</simultáneo> <simultáneo> es que esa</simultáneo> persona claro / estará trabajando y ganando su dinero / y sin embargo cuando no la	España	Transcripción | Audio
17	ALCA_M11_041	<alargarmiento/> uno o dos hijos / porque luego también influye a lo mejor <vacilación/> una persona que tenga cinco o seis hijos / desde luego no <simultáneo> te puedes</simultáneo> <simultáneo>¡hombre!</simultáneo>	España	Transcripción | Audio
18	ALCA_M12_023	tampoco se está muy seguro es que no sé / te puedes encontrar a cualquier persona por ahí que <alargarmiento/> puede parecer normal y <alargarmiento/> claro no te puedes fiar ¿	España	Transcripción | Audio
19	ALCA_M12_023	tenía el de la moto // también no llevábamos casco era una moto de una persona y <alargarmiento/> / pero no tenía que habe <alargarmiento/> r o sea tenía <vacilación/> la	España	Transcripción | Audio
20	ALCA_M13_005	/ sí / ¿y tú cómo lo haces? // normalmente yo a las persona <alargarmiento/> s mayores / mmm / y por ejemplo del trabajo también / sí /	España	Transcripción | Audio

a Enlace para descargar los resultados en un fichero CSV: texto en Unicode UTF-8, con columnas separadas por tabuladores, la columna correspondiente al término buscado incluye dicho término, la <etiqueta gramatical> y el [lema]

personas <NCFP000> [persona]
hazlo <VMM02S0+PP3CNA00> [hacer+lo]
tenía <VMII3S0> [tener] miedo <NCMS000> [miedo]

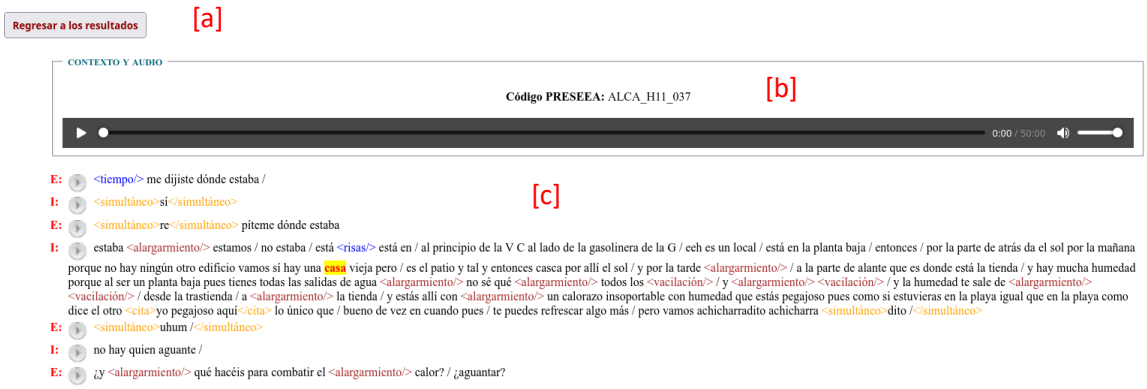
b Resultados de la consulta

- Consulta efectuada: consulta + filtros utilizados, p. ej.: [(word=persona%c)] :: (match.text_ciudad = "Alcalá de Henares" | match.text_ciudad = "Barcelona" | match.text_ciudad = "Barranquilla") & (match.text_sexo = "M") & (match.text_nivel = "3") & match.turno_inf = "I.*" within turno
- Número total de ejemplos en número de textos
- Frecuencia normalizada sobre el subcorpus seleccionado: casos por cada 100.000 palabras
- Indicador de resultados mostrados: permite avanzar o retroceder

c Los resultados se presentan en columnas

- Número secuencial del resultado
- Clave: el código PRESEEA de la entrevista que contiene el ejemplo.
- Concordancia (KWIC): Contexto anterior + Búsqueda + Contexto posterior.

Al hacer clic en cada elemento de la columna **Clave** se muestra el término buscado dentro de su contexto ampliado correspondiente, distinguiéndose los siguientes elementos:



a Botón para regresar a los resultados de la consulta

b Clave de la entrevista y reproductor de audio

c Contexto extendido: se muestra el fragmento con el término consultado resaltado: casa. Se muestran 3 turnos de habla anteriores y posteriores al turno donde ocurre el término buscado. Cada turno de habla puede reproducirse presionando el icono de audio que antecede a la transcripción de dicho turno.

B. Tipos de consulta

La interfaz está diseñada de manera que no es necesario estar familiarizado con la sintaxis de consulta que utiliza el corpus PRESEEA, pudiéndose escribir una consulta básica (con comodines o sin ellos) en la casilla de entrada de datos [a] del **buscador** ①, por ejemplo la expresión `mezquin.+`, y presionar directamente el botón **Buscar** [c].

También es posible generar de forma semi-automática un elevado número de consultas mediante el **generador de consultas** ②.

a. forma ortográfica

`(word='perro'%c)`

1 Introducir término en la caja “Forma ortográfica”
2 Hacer clic en el botón **Crear consulta ⇑**
3 Hacer clic en el botón **Buscar**

b. lema

`[(lemma='perro'%c)]`

1 Introducir término en la caja “Lema”
2 Hacer clic en el botón **Crear consulta ⇑**
3 Hacer clic en el botón **Buscar**

c. etiqueta gramatical

```
[(pos='AQ.MS.'%c)]
```

1 Introducir etiqueta gramatical en la caja “Etiqueta gramatical”
2 Si se desconoce la etiqueta, hacer clic en **?** para mostrar la ayuda para definir la etiqueta. Una vez definida, hacer clic en “finalizar”
3 Hacer clic en el botón **Crear consulta ⇑**
4 Hacer clic en el botón **Buscar**

d. término 1 o término 2

```
[(lemma='perro'%c) | (lemma='gato'%c)]
```

1 Introducir término en la caja correspondiente
2 Hacer clic en **O** para añadir el segundo término
3 Introducir segundo término en la caja correspondiente
4 Hacer clic en el botón **Crear consulta ⇑**
5 Hacer clic en el botón **Buscar**

e. término 1 y término 2

```
[(lemma='casa'%c)] [(lemma='grande'%c)]
```

1 Introducir término en la caja correspondiente [término 1]
2 Hacer clic en **+** para añadir el segundo término
3 Introducir término en la caja correspondiente [término 2]
4 Hacer clic en el botón **Crear consulta ⇑**
5 Hacer clic en el botón **Buscar**

f. término 1 y término 2 con operador de repetición

```
[(lemma='perro'%c)] []{0,2} [(lemma='gato'%c)]
```

1 Introducir término en la caja correspondiente [término 1]
2 Hacer clic en **↔** para añadir el operador de repetición
3 Modificar el operador de repetición []{0,2} (el término anterior puede aparecer repetido entre cero y dos veces)
4 Hacer clic en **+** para añadir el segundo término
5 Introducir término en la caja correspondiente [término 2]
6 Hacer clic en el botón **Crear consulta ⇑**
7 Hacer clic en el botón **Buscar**

g. término 1a o 1b y término 2a o 2b o 2c o 2d con operador de repetición

```
[(lemma='haber'%c) | (lemma='tener'%c)] []{0,3} [(lemma='miedo'%c) |
(lemma='vergüenza'%c) | (lemma='frío'%c) | (lemma='hambre'%c)]
```

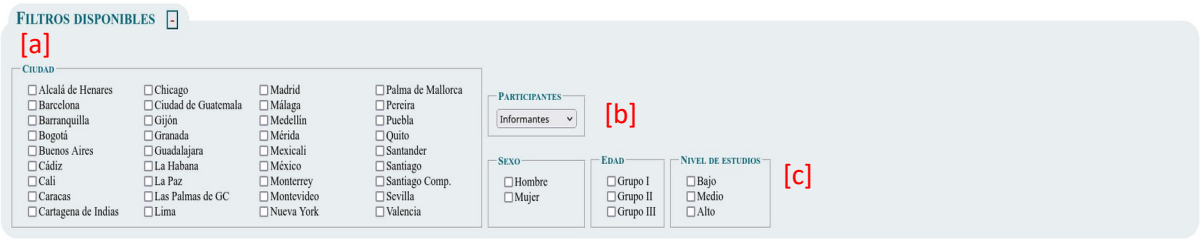
1 Introducir término en la caja correspondiente: *haber*
2 Hacer clic en **O** para añadir el segundo término
3 Introducir segundo término en la caja correspondiente: *tener*
4 Hacer clic en **↔** para añadir el operador de repetición
5 Modificar el operador de repetición []{0,3} (el término anterior puede aparecer repetido entre cero y tres veces)

6 Introducir término en la caja correspondiente: *miedo*
7 Hacer clic en **O** para añadir el segundo término
8 Introducir término en la caja correspondiente: *vergüenza*
9 Hacer clic en **O** para añadir el tercer término
10 Introducir término en la caja correspondiente: *frío*
11 Hacer clic en **O** para añadir el cuarto término
12 Introducir término en la caja correspondiente: *hambre*
13 Hacer clic en el botón **Crear consulta ⇑**
14 Hacer clic en el botón **Buscar**

Sin embargo, algunas consultas más complejas requieren que los datos sean introducidos directamente en la **casilla de entrada de datos.** En el ***Apéndice A*** se ofrece una explicación detallada de la sintaxis CQP y una serie de ejemplos concretos de diferentes tipos de consulta.

C. Filtrado

Una vez ejecutada una consulta es posible filtrar los resultados utilizando los "Filtros disponibles" sin necesidad de volver a recrear la consulta.



- [a] Ciudad: casillas de selección múltiples
- [b] Participantes: menú desplegable (Informantes, entrevistadores o ambos)
- [c] Sexo, edad y nivel de estudios: casillas de selección múltiples

D. Cómo citar el corpus PRESEEA

Los materiales del Corpus PRESEEA pueden ser consultados sin coste alguno. Su uso ha de estar destinado exclusivamente a la investigación, quedando prohibido el empleo de cualquier material PRESEEA con fines comerciales. No olvide que la utilización de los materiales obliga a citar siempre su procedencia, que ha de hacerse de la siguiente manera:

> PRESEEA (2014-): *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá. [http://preseea.uah.es]. Consultado: […]

Para la consulta de otros materiales, los interesados han de ponerse en contacto con los coordinadores de cada equipo, cuyas direcciones aparecen en la sección "Contacto".

Para permitir que otros investigadores puedan comprobar los resultados obtenidos, se recomienda incluir la consulta efectuada tal como aparece en los **resultados** ④, incluyendo no solo el término o términos o expresión buscados, sino también los filtros empleados:

- Q = `[(lemma='perro'%cd)]` within turno
- Q = `[(lemma='aceite'%cd)]` :: (match.text_ciudad = "Barcelona") & (match.text_sexo = "M") & (match.text_edad = "2") & (match.text_nivel = "2") & match.turno_inf = "I.*" within turno

Apéndice A: Breve introducción a la sintaxis CQP (*Corpus Query Processor*)

El corpus PRESEEA utiliza para las consultas la colección de herramientas de código abierto IMS Open Corpus Workbench (CWB)[1]. Su componente central es el procesador de consultas CQP (*Corpus Query Processor*) que permite realizar consultas utilizando la coincidencia de patrones con expresiones regulares. A continuación se ofrece una breve introducción a la sintaxis CQP[2].

Una de las utilidades de los corpus lematizados y etiquetados morfológicamente, como el PRESEEA, es el hecho de que se pueden efectuar consultas por atributos, es decir, por lema, palabra o categoría morfológica ("PoS" *Part of Speech*).

Lema	`[(lemma='perro')]` → *perro, perra, perros, perrita*, etc.
Palabra	`[(word='perro')]` → *perro*
PoS	`[(pos='AQ.MP.')]` → todos los *adjetivos calificativos masculino plural*

1 Búsqueda por lemas

Al efectuar una búsqueda por lemas el resultado serán todas las formas de ese lema en el corpus. Para buscar por lema, debemos utilizar la siguiente expresión `[(lemma='reina')]`. Si lo que se desea son cadenas de lemas, simplemente debemos repetir la expresión anterior tantas veces como lemas, el espacio en blanco entre ellas es opcional: `[(lemma='reina')] [(lemma='don')]`. *NOTA: No hay que dejar espacios en blanco entre las comillas y el término buscado, pues de esta forma el buscador no devolverá ningún resultado*.

Es preciso recordar que, en las consultas por lema, las mayúsculas y minúsculas son significativas y la consulta `[(lemma='alfonso')]` no devuelve ningún resultado, siendo necesario escribir el lema en mayúsculas `[(lemma='Alfonso')]` o utilizar el operador `%c` en la consulta `[(lemma='alfonso'%c)]`. Lo mismo ocurre con los diacríticos y la consulta `[(lemma='Gutierrez')]` no devuelve ningún resultado, siendo necesario poner la tilde `[(lemma='Gutiérrez')]` o utilizar el operador `%d` en la consulta `[(lemma='Gutiérrez'%d)]` para encontrar ejemplos de palabras correspondientes al lema *Gutiérrez*.

Finalmente hay que señalar que los operadores `%c` y `%d` pueden combinarse en la misma consulta: `[(lemma='gutierrez'%cd)]`

2 Búsqueda por palabras

Para buscar palabras, debemos utilizar la siguiente expresión `[(word='casa')]`. Si lo que se desea son cadenas de palabras, simplemente debemos repetir la expresión anterior tantas veces como palabras, el espacio en blanco entre ellas es opcional: `[(word='casa')] [(word='grande')]`. *NOTA: No hay que dejar espacios en blanco entre las comillas y el término buscado, pues de esta forma el buscador no devolverá ningún resultado.*

[1] http://cwb.sourceforge.net/

[2] Para mayor información sobre el lenguaje de consulta CQP consúltese Evert, Stefan, et al. (2019) *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial. CWB Version 3.4.16*. (http://cwb.sourceforge.net/files/CQP_Tutorial.pdf)

Es preciso recordar que, en las consultas por palabra, las mayúsculas y minúsculas son significativas y la consulta `[(word='hombre')]` solo incluye ejemplos sin mayúscula (*hombre*), siendo necesario utilizar el operador `%c` para encontrar también ejemplos con mayúscula, `[(word='hombre'%c)]` → *hombre*, *Hombre*, *HOmbre*, etc.

Lo mismo ocurre con los diacríticos y la consulta `[(word='méxi.*')]` solo incluye ejemplos de palabras *é* por (*méxico*), siendo necesario utilizar el operador `%d` para encontrar también ejemplos con *e* sin tilde `[(word='mexico.*'%d)]` → *méxico*, *mexico*, *mexicanas*, etc.

Los operadores `%c` y `%d` pueden combinarse en la misma consulta: `[(word='México'%cd)]` → *México*, *Mexico*, *Mexicanos, mexicanas,* etc.

3 Búsqueda por PoS

Para buscar categorías morfológicas (PoS), debemos utilizar la siguiente expresión `[(pos='NCFS000'%c)]`. *NOTA: La etiqueta morfológica debe ser siempre escrita en mayúscula*. Si lo que se desea son cadenas de palabras, simplemente debemos repetir la expresión anterior tantas veces como palabras, el espacio en blanco entre ellas es opcional: `[(pos='NCFS000'%c)] [(pos='AQ0FS0'%c)]`. *NOTA: No hay que dejar espacios en blanco entre las comillas y el término buscado, pues de esta forma el buscador no devolverá ningún resultado. NOTA: En el* ***Apéndice C*** *se ofrece la lista completa de las etiquetas EAGLES utilizadas en el corpus PRESEEA.*

4 Búsqueda por frecuencia

Es posible buscar palabras o expresiones por la frecuencia (exacta, mínima o máxima) con la que aparecen en el corpus. Para encontrar todas las palabras que comienzan por 'p' en el corpus y que solo ocurren una vez debemos utilizar la expresión `[(word='p.+'%c) & (f(word) = 1)]`. Para encontrar todos los lemas que ocurren menos de 5 veces debemos utilizar la expresión `[(lemma='.+'%c) & (f(lemma) < 5)]`. Para encontrar todas las palabras que terminan en 'orio' en el corpus y que ocurren más de 2000 veces debemos utilizar la expresión `[(word='.+orio'%c) & (f(word) > 2000)]`. Finalmente, puede establecerse un rango de frecuencia mínima y máxima con la expresión `[(word='p.+'%c) & (f(word) > 100 & f(word) < 1000)]`.

5 Caracteres comodín y operadores

Supongamos ahora que queremos encontrar todas las formas del sustantivo "planta"; para ello podríamos hacer las siguientes dos consultas: `[(word='planta'%c)]` y `[(word='plantas'%c)]`, sin embargo, además de no tener los resultados en una única consulta, vemos que estos incluyen tanto la forma verbal como el sustantivo. Para evitar esto podemos utilizar caracteres comodín y operadores y efectuar una búsqueda combinada para especificar tanto formas en singular y plural como la categoría morfológica que queremos: `[(word='plantas?' & pos='NC.*'%c)]`, y lo que ahora buscamos son las palabras *planta* o *plantas* cuando son un sustantivo (con el comodín `?` indicamos que el carácter que le

precede puede aparecer cero veces o una vez, y con el operador `.*` indicamos que la etiqueta PoS debe empezar por NC). A continuación presentamos los caracteres comodín y los operadores más habituales:

`.` equivale a un único carácter: `[(word='ca.a'%c)]` → *cada*, *cala*, *cama*, *cara*, *casa*, …; `[(word='ca..a'%c)]` → *cabra*, *caída*, *carta*, *causa*, …

`.*` equivale a cero o más caracteres: `[(word='amarga.*'%c)]` → *amarga*, *amargad*, *amargado*, *amargamente*, …

`.+` equivale a uno o más caracteres: `[(word='amarga.+'%c)]` → *amargad*, *amargado*, *amargamente*, … pero no *amarga*

`..+` equivale a tres o más caracteres: `[(word='amarga..+'%c)]` → *amargados*, *amargamente* … pero no *amarga*, *amargar*, *amargado*

`?` el caracter que le precede puede aparecer cero veces o una vez: `[(word='res?pond?er'%c)]` → *responder* y *reponer*

`*` el carácter que le precede puede aparecer cero o más veces: `[(word='cas*as*'%c)]` → *casa*, *casas*, *cassa*, *cassas*

`+` el carácter que le precede puede aparecer una o más veces: `[(word='car+o'%c)]` → *caro* y *carro*

`{n}` el carácter o la expresión que le precede puede aparecer tantas veces como indique n (nº de ocurrencias): `[(word='.*o{2}.*'%c)]` → para todas las palabras que contienen dos oes seguidas; `[(pos='NP.*'%c)]{2}` → para obtener todas las ocurrencias de dos nombres propios seguidos: *Juan Pedro*, *Luis García*, etc.

`{m,n}` operador de repetición (m = número mínimo, n = número máximo de palabras): `[(pos='DA.*'%c)] [(pos='A.*'%c)]{1,2} [(lemma='hombre'%c) | (lemma='mujer'%c)]` → para obtener todas las ocurrencias de un artículo definido seguido por uno o dos adjetivos, más cualquiera de las formas de los lemas hombre o mujer.

`[]` los corchetes, sin espacio entre ellos, equivalen a una única palabra: `[(word='yo'%c)] [] [(word='doy'%c)]` → para encontrar las palabras que aparecen entre *yo* y *doy*; `[(lemma='dar'%c)] [] [(pos='SP.*'%c)] []` → para encontrar todas las palabras que ocurren en una cadena de palabras.

`[ ]` cualquiera de los caracteres dentro de los corchetes puede aparecer como mucho una vez: `[(word='pro[bv]ar'%c)]` → *probar*, *provar*.

`[ ]*` cualquiera de los caracteres dentro de los corchetes puede aparecer cero o más veces: `[(word='médic[oa][s]*'%c)]` → *médico*, *médicos, médica*, etc.

`[^ ]` cualquier carácter, excepto los caracteres dentro de los corchetes, puede aparecer: `[(word='a[^mt]ar'%c)]` → *asar*, *azar* pero no *amar* o *atar*

`!` operador de negación, `[(pos!='V.*'%c)]` → para encontrar todas las formas de las categorías gramaticales presentes en una obra, autor, etc., excepto las verbales

Diferentes caracteres comodín y operadores pueden combinarse para efectuar consultas más complejas. Por ejemplo:

- `[(word='h?a[bv]er'%c)]` → para buscar todas las variantes ortográficas (correctas o no) del infinitivo *haber*.

6 Búsquedas combinadas

Las búsquedas combinadas son aquellas que agrupan varios elementos en la misma consulta. Existen cuatro tipos básicos de búsqueda combinada:

6.1 Secuencial o yuxtapuesta

En este tipo de consulta combinamos dos o más elementos sin ninguna relación entre ellos aparte de la meramente secuencial. `[(lemma='hablar'%c)] [(word='mucho'%c)]` → para encontrar la secuencia *habla+mucho*; `[(lemma='hablar'%c)] [(pos='NP.*'%c)]` → para encontrar todas las formas de *hablar* seguidas por antropónimo.

6.2 Copulativa (`&`)

En este tipo de consulta buscamos los registros que **incluyan todos** los términos unidos mediante el operador `&`. `[(lemma='traer' & pos='V.II.*'%c)]` → para encontrar las formas de pretérito imperfecto de indicativo del verbo *traer*; `[(word='habla') & (lemma='hablar')]` → para encontrar la palabra *habla* cuando pertenece al lema *hablar*; `[(pos='V.*'%c) & (lemma='pro.*'%c)]` → para encontrar todos los verbos que comienzan por *pro-*.

6.3 Disyuntiva (`|`)

En este tipo de consulta buscamos los registros que **incluyan cualquiera** de los términos unidos mediante el operador `|`. `[(word='alto'%c) | (word='bajo'%c)]` → para encontrar las palabras *alto* o *bajo*.

6.4 Negativa (`&!`)

En este tipo de consulta buscamos los registros que **excluyan** el segundo término de los dos que aparecen unidos por el operador `&!`. `[(lemma="sierra"%c) &! (pos="V.*"%c)]` → para encontrar todas las formas del lema *sierra* excepto cuando su categoría morfológica es verbo. `[(word='a.*'%c) &! (pos='SP.*'%c)]` → para buscar todas las palabras que comienzan por *a* excepto cuando son preposiciones.

NOTA: El orden de los términos en la consulta no es arbitrario. Así `[(lemma='f.*' & pos='NP000G0'%c)]` → para encontrar todas las formas de los lemas que comienzan por *f* y son topónimos, mientras que `[(pos='NP000G0' & lemma='f.*'%c)]` → para buscar todas las palabras etiquetadas como topónimos que comienzan por *f*.

7 Búsquedas complejas

Las búsquedas complejas son aquellas que agrupan varias búsquedas combinadas en la misma consulta. A continuación se ofrecen algunos ejemplos:

- `[(lemma='traer'%c) & (pos='V.II.*' | pos='V.SI.*'%c)]` → para encontrar tanto las formas de pretérito imperfecto de indicativo como las de subjuntivo del verbo *traer*; esta búsqueda puede también escribirse así: `[(lemma='traer'%c) & (pos='V.[IS]I.*'%c)]`
- `[(lemma='traer'%c) &! (pos='V.[IS]I.*'%c)]` → para encontrar todas las formas del verbo *traer*, excepto las de pretérito imperfecto de indicativo y las de pretérito imperfecto de subjuntivo.
- `[(word='a.*'%c) &! (pos='SP.*' | pos='V.*'%c)]` → para encontrar todas las ocurrencias de *a*, excepto cuando se trata de una preposición o un verbo.
- `[(pos='NC.*'%c) & (word='[aeiou].*'%c)] [(pos='SP.*'%c) & (word='de'%c)] [(pos='NC.*'%c) & (word='[aeiou].*'%c)]` → para buscar dos sustantivos que comienzan con una vocal y están relacionados entre sí por la preposición “de”.
- `[!(lemma='haber'%c) | (lemma='tener'%c)] [(lemma='bien'%c)]` → *bien* precedido por cualquier palabra, excepto las correspondientes a los lemas *haber* o *tener*.
- `[(lemma='bien'%c)] [!(lemma='haber'%c) | (lemma='tener'%c)]` → *bien* seguido por cualquier palabra, excepto las correspondientes a los lemas *haber* o *tener*.

8 Ejemplos de consultas

Para mostrar las posibilidades del corpus PRESEEA, ofrecemos a continuación una serie de ejemplos concretos de consultas con los filtros utilizados. Debe recordarse que es posible descargar los resultados en un fichero CSV para su posterior manipulación con otros programas (Excel, R, etc.)

8.1 Consultas por marca estructural

Los textos incluidos en el corpus PRESEEA contienen una serie de marcas estructurales que reflejan la disposición del texto en la página. Esta consulta devuelve la primera palabra de cada turno:

```
<turno>[(word='.*'%c)]
```
within turno

Esta consulta devuelve los turnos de habla que comienzan con el marcador discursivo “pues”:

```
<turno>[(word='pues'%c)]
```
within turno

8.2 Extracción de todas las formas del imperfecto de subjuntivo

En la forma más sencilla de esta consulta podemos usar la etiqueta morfológica:

```
[(pos='V.SI.*'%c)]
```
within turno

para obtener una lista de todas las formas del imperfecto de subjuntivo, sin embargo podemos afinar aún más nuestros resultados con dos consultas que nos permitan sacar una lista de las formas tipo *amara* y otra de las formas tipo *amase*:

```
[(word='.*[ea]ra.*' & pos='V.SI.*'%c)]
```
within turno
```
[(word='.*[ea]se.*' & pos='V.SI.*'%c)]
```
within turno

8.3 Extracción de colocaciones del tipo “venirle en voluntad”

Esta consulta busca 1) cualquier forma de *venir* seguida por 2) cualquier forma de *voluntad*, encontrándose 1) y 2) separadas por una palabra (*vino a voluntad*, *venían con voluntad*, etc.):

```
[(lemma='venir'%c)] [] [(lemma='voluntad'%c)]
```
within turno

Si utilizamos el operador `?` la consulta busca 1) cualquier forma de *venir* seguida por 2) cualquier forma de *voluntad*, encontrándose 1) y 2) separadas por cero o una palabra (*venga voluntad*, *viene en voluntad*, etc.):

```
[(lemma='venir'%c)] []? [(lemma='voluntad'%c)]
```
within turno

Podemos también delimitar el número de palabras que separan 1) y 2). Para ello podemos repetir el operador `[]` tantas veces como palabras separan 1) y 2) (*vino hoy a voluntad*, *venir de buena voluntad*, etc):

```
[(lemma='venir'%c)] [] [] [(lemma='voluntad'%c)]
```
within turno

o bien podemos utilizar el operador de repetición `{m,m}` para determinar el número mínimo y máximo de palabras que pueden separar 1) y 2) (*vino voluntad*, *venía con voluntad*, *vinieron con mucha voluntad*, *venía de muy buena voluntad*):

```
[(lemma='venir'%c)] []{0,3} [(lemma='voluntad'%c)]
```
within turno

Finalmente, podemos encontrar colocaciones similares aumentando el número de términos en 1) y 2):

```
[(lemma='venir'%c) | (lemma='entrar'%c) | (lemma='caer'%c)] []{0,3}
  [(lemma='deseo'%c) | (lemma='propósito'%c) | (lemma='voluntad'%c) |
  (lemma='pensamiento'%c)] within turno
```

8.4 Extracción de colocaciones del tipo "verbo+pronombre_sujeto / pronombre_sujeto+verbo"

La siguiente consulta busca todas las formas de 1ª persona del presente de indicativo de los verbos *ser*, *dar*, *ir*, *estar*, precedidas por cualquier forma del pronombre *yo*:

```
[(pos='PP1CSN00'%c)] [(lemma='dar' & pos='V.IP1S.'%c) | (lemma='ser' &
  pos='V.IP1S.'%c) | (lemma='estar' & pos='V.IP1S.'%c) | (lemma='ir' &
  pos='V.IP1S.'%c)] within turno
```

Mientras que la siguiente consulta busca todas las formas de 1ª persona del presente de indicativo de los verbos *ser*, *dar*, *ir*, *estar*, seguidas por cualquier forma del pronombre *yo*:

```
[(lemma='dar' & pos='V.IP1S.'%c) | (lemma='ser' & pos='V.IP1S.'%c) |
  (lemma='estar' & pos='V.IP1S.'%c) | (lemma='ir' & pos='V.IP1S.'%c)]
  [(pos='PP1CSN00'%c)] within turno
```

Es posible combinar ambas consultas y buscar simultáneamente todas las formas de 1ª persona del presente de indicativo de los verbos *ser*, *dar*, *ir*, *estar*, precedidas o seguidas por cualquier forma del pronombre *yo*:

```
[(pos='PP1CSN00'%c)] [(lemma='dar' & pos='V.IP1S.'%c) | (lemma='ser' &
  pos='V.IP1S.'%c) | (lemma='estar' & pos='V.IP1S.'%c) | (lemma='ir' &
  pos='V.IP1S.'%c)] | [(lemma='dar' & pos='V.IP1S.'%c) | (lemma='ser' &
  pos='V.IP1S.'%c) | (lemma='estar' & pos='V.IP1S.'%c) | (lemma='ir' &
  pos='V.IP1S.'%c)] [(pos='PP1CSN00'%c)] within turno
```

8.5 Extracción de casos de desdoblamiento de género "*niños y niñas*", "*los niños y las niñas*"

Las siguientes consultas nos permiten extraer todos los pares de sustantivos masculinos y femeninos unidos por cualquier forma la conjunción copulativa "y" (*y*, *e*), con o sin artículo definido delante del sustantivo.

```
[(pos='NCM.*'%c)] [(lemma='y'%c)] [(pos='NCF.*'%c)] within turno sort by word
[(pos='DA.M.*'%c)] [(pos='NCM.*'%c)] [(lemma='y'%c)] [(pos='DA.F.*'%c)]
  [(pos='NCF.*'%c)] within turno
```

8.6 Extracción de verbos+pronombre_enclítico (*quiéranlos*, *tenerlos*, etc.)

Para buscar ciertas formas aglutinadas o los verbos+pronombre enclítico es preciso utilizar o bien la almohadilla (#) o bien el punto medio (· en los teclados que lo incluyen). Esta consulta permite encontrar todas las formas aglutinadas que no sean verbos

```
[(lemma='.+#.+' & pos!='V.+'%c)] within turno
```

y esta otra todos los verbos con un pronombre_enclítico

```
[(pos='V.+#.+'%c)] within turno
```

8.7 Extracción de colocaciones del tipo "sustantivo precedido por cualquier palabra excepto las de una categoría gramatical concreta".

En este caso queremos extraer las colocaciones del sustantivo *peso* precedido por cualquier palabra excepto artículo. Podemos utilizar una de estas dos expresiones para encontrar las formas del sustantivo:

```
[(word='pesos*' & pos='NC.*'%c)] within turno
[(lemma='peso')] within turno
```

Para eliminar todos aquellos casos en los que *peso* está precedido por un artículo, añadimos la expresión `[(pos!='DA.*'%c)]` a nuestra consulta:

```
[(pos!='DA.*'%c)] [(word='pesos*' & pos='NC.*'%c)] within turno
```

de manera que se buscan combinaciones de dos palabras, donde la primera de ellas NO sea un artículo y la segunda sea *peso* SOLO cuando se trata de un nombre común.

Para excluir los artículos y las formas aglutinadas con artículos debemos utilizar la siguiente búsqueda:

```
[(pos='.*'%c) &! (pos='DA.*'%c) &! (pos='SPS00#DA.*'%c)] [(word='pesos*'
& pos='NC.*'%c)] within turno
```

en la que se buscan combinaciones de dos palabras, donde la primera de ellas tenga cualquier etiqueta morfológica, excepto artículo o la forma aglutinada preposición+artículo.

Apéndice B: Etiquetas EAGLES

El conjunto de etiquetas morfológicas utilizadas en el corpus PRESEEA se basa en las etiquetas propuestas por el grupo *Expert Advisory Group on Language Engineering Standards* (EAGLES)[3]. A continuación se presentan cada una de las etiquetas definidas en una tabla y el conjunto de etiquetas utilizadas en el PRESEEA. Para cada categoría se presentan los atributos, valores y códigos que puede tomar, así como algunos ejemplos relevantes. Las tablas en la que se presentan las etiquetas tienen el siguiente aspecto:

Categoría			
Posición	*Atributo*	*Valor*	*Código*
columna 1	columna 2	columna 3	columna 4

En la *columna 1* encontramos un número que hace referencia al orden y posición en que aparecen los atributos. La *columna 2* hace referencia a los atributos, el número de los cuales varía dependiendo de la categoría. En la *columna 3* encontramos los valores que puede tomar cada atributo y, finalmente, la *columna 4* representa los códigos que se han establecido para su representación. Las etiquetas en sí sólo son los códigos (columna 4) y se sabe a qué atributo pertenecen por la posición (columna 1) en la que se encuentran. Así NCMS000, correspondiente a *nombre* (N), *común* (C), *masculino* (M), *singular* (S), solo tiene cuatro códigos, mientras que VMIS2P0, *verbo* (V), *principal* (M), *pretérito perfecto* (I), *simple* (S), *2 persona* (2), *plural* (P), tiene seis.

Adjetivos			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
		Ordinal	O
3	Grado	Aumentativo	A
		Diminutivo	D
		Comparativo	C
		Superlativo	S
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Función	-	0
		Participio	P

[3] http://www.ilc.cnr.it/EAGLES/browse.html

Adverbios			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Adverbio	R
2	Tipo	General	G
		Negativo	N
		-mente	M

Conjunciones			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Conjunción	C
2	Tipo	Coordinada	C
		Subordinada	S

Determinantes			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Determinante	D
2	Tipo	Demostrativo	D
		Posesivo	P
		Interrogativo	T
		Exclamativo	E
		Indefinido	I
		Artículo	A
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Poseedor	Singular	S
		Plural	P

Interjecciones			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Interjección	I

Nombres			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5-6	Clasificación semántica	Persona	P0
		Lugar	G0
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

Numerales			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Cifra	Z

Preposiciones			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Adposición	S
2	Tipo	Preposición	P
3	Forma	Simple	S
		Contraída	C
4	Género	Masculino	M
5	Número	Singular	S

Pronombres			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Pronombre	P
2	Tipo	Personal	P
		Demostrativo	D
		Posesivo	X
		Indefinido	I
		Interrogativo	T
		Relativo	R
		Exclamativo	E
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Impersonal Invariable	N
6	Caso	Nominativo	N
		Acusativo	A
		Dativo	D
		Oblicuo	O
7	Poseedor	Singular	S
		Plural	P
8	Politeness	Polite	P

Verbos			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

Signos de puntuación			
Posición	*Atributo*	*Valor*	*Código*
1	Categoría	Puntuación	F

Apéndice C: Marcas estructurales

Las etiquetas utilizadas en las transcripciones para indicar la estructura del texto o la presencia de diferentes elementos en cada grabación son las siguientes.

Etiqueta PRESEEA[4]	Descripción
<text>	*Encabeza cada transcripción del corpus*
<text_id>	*Identificador de la entrevista*
<turno>	*Turno de habla*
<lengua>	*Otras lenguas distintas al español*
<ruido>	*Ruidos recogidos en la grabación*
/	*Pausa*
//	*Pausa larga*
<risas>	*Risa sin contenido lingüístico*
<entre_risas>	*Texto enunciado entre risas*
<enfasis>	*Énfasis en la pronuncación*
<simultaneo>	*Conversación simultánea*
<vacilación>	*Vacilación en la emisión*
<siglas>	*Siglas*
<cita>	*Cita textual del informante*
<alargamiento>	*Alargamiento de palabras en la emisión*
<palabra_cortada>	*Palabra cortada*
<solapamiento>	*Emisiones solapadas entre dos o más hablantes*
<tiempo>	*Marca de tiempo de cada turno*

[4] Cada una de estas etiquetas posee su correspondiente etiqueta de cierre: </turno>, </lengua>, etc.