Guía básica de uso

Esta es una guía básica para realizar búsquedas en los corpus del Instituto de Investigaciones Lingüísticas. El software que utiliza el sistema de corpus es ANNIS (ANNotation of Information Structure). Para hacer búsquedas, el programa utiliza el lenguaje de búsquedas de ANNIS (*ANNIS query language* o AQL). Para una versión más detallada del AQL consultar este sitio. Asimismo, esta documentación es útil para realizar búsquedas y explotar otras características más avanzadas del programa.

Índice:

1. Búsquedas sencillas

Se puede buscar una palabra en el corpus sencillamente colocándola entre comillas:

“juez”

Es importante notar que la búsqueda es sensible a mayúsculas, por lo que se puede emplear la siguiente fórmula si se quieren recuperar casos con o sin mayúscula:

“Juez” | “juez”

Se pueden buscar secuencias de palabras con el operador “.” . La siguiente búsqueda devolvería “El juez” si esto apareciera en el corpus.

“El” . “juez”

2. Búsquedas con anotación

1. Clase de palabra

El corpus está anotado con información de clase de palabra (PoS). Se puede buscar una clase de palabra específica. Por ejemplo, la siguiente búsqueda devuelve todos los sustantivos del corpus:

clase=”NOUN”

Asimismo, las búsquedas de palabras específicas se pueden delimitar con una clase de palabra. Por ejemplo, si se desea buscar los casos en los que aparece la palabra “azul” como sustantivo, se puede utilizar la siguiente búsqueda :

“azul” _=_ clase=”NOUN”

El operador _=_ indica que lo que está a su izquierda y lo que está a su derecha aplican al mismo token. En esta caso, esta búsqueda encuentra tokens para los cuales “azul” y clase=”NOUN” son ciertos.

La siguiente tabla resume las opciones para POS

Argumento (POS=”...”)	Significado
NOUN	Sustantivos
VERB	Verbos
AUX	Verbo auxiliar
ADP	Adposición (pre- y posposiciones)
ADJ	Adjetivo
ADV	Adverbio
PRON	Pronombre
DET	Determinante
PROPN	Nombre propio
CONJ	Conjunción
SCONJ	Conjunción subordinante
PUNCT	Puntuación

2. Lema

El corpus fue anotado con lemas para cada token. El lema es la forma base de una palabra (por ejemplo, el lema de “somos” es “ser”).

Para buscar una palabra por su lema, es tan sencillo como hacer la siguiente búsqueda:

lema=”ser”

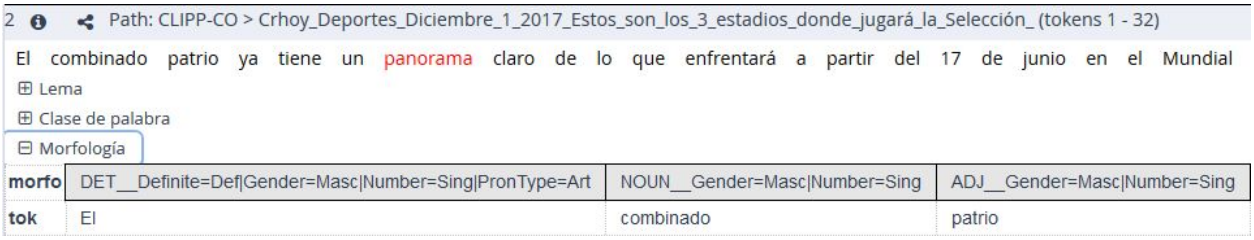
Al igual que con clase de palabras, se puede delimitar una búsqueda con otros tipos de anotación o con el contenido textual del token:

“fuimos” _=_ lema=”ser”

Esta búsqueda encuentra tokens para los que aplique “fuimos” y cuyo lema es “ser” (esto excluye el “fuimos” cuyo lema es “ir”).

3. Información morfológica

El corpus fue anotada con información morfológica como género, número, persona, etc:



Para buscar rasgos morfológicos particulares, es necesario utilizar expresiones regulares. Consultar la sección 6 y esta página para una introducción al uso de expresiones regulares.

La siguiente expresión regular busca tokens en singular

morfo=/.*Sing.*/

Porque concuerda con strings de anotación como los siguientes:

Token	morfo
“tiene”	VERB__Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
“claro”	ADJ__Gender=Masc|Number=Sing
“la”	DET__Definite=Def|Gender=Fem|Number=Sing|PronType=Art

La siguiente expresión regular concuerda únicamente con tokens que son femeninos y singulares

morfo=/.*Fem.*Sing.*/

tales como:

Token	morfo
“ciudad”	NOUN__Gender=Fem|Number=Sing
“una”	DET__Definite=Ind|Gender=Fem|Number=Sing|PronType=Art
“plagada”	ADJ__Gender=Fem|Number=Sing|VerbForm=Part

NOTA: Si se desea limitar la búsqueda con más de un rasgos morfológico, es importante ver cuál aparece antes en la anotación, y utilizar este mismo orden en la expresión regular. Por ejemplo, la siguiente búsqueda no devuelve resultados, pues el género se marca antes que el número en este esquema:

morfo=/.*Sing.*Fem.*/

4. Resumen de anotaciones

Anotación	Descripción	Valores posibles
clase	Clase de palabra	Tags Universales
lema	Forma base de una palabra	
morfo	Información morfológica (género, número, persona, etc.)	Link

3. Búsquedas con metadatos

Los textos de CLIPP-CO provienen de diferentes géneros periodísticos y de diferentes fuentes. Una búsqueda se puede especificar con alguno de estos metadatos de la siguiente forma:

```
"juez" & meta::Seccion="Nacionales"
```

La siguiente tabla resume argumentos posibles para los metadatos y sus valores. Se debe tomar en cuenta que estos argumentos y valores deben ser escritos tal cual aparecen en la tabla.

Metadato	Valores posibles
Fuente	Crhoy, Nacion, Extra
Seccion	Opinion, Deportes, Nacionales, Sucesos

4. Ejemplo de una búsqueda compleja

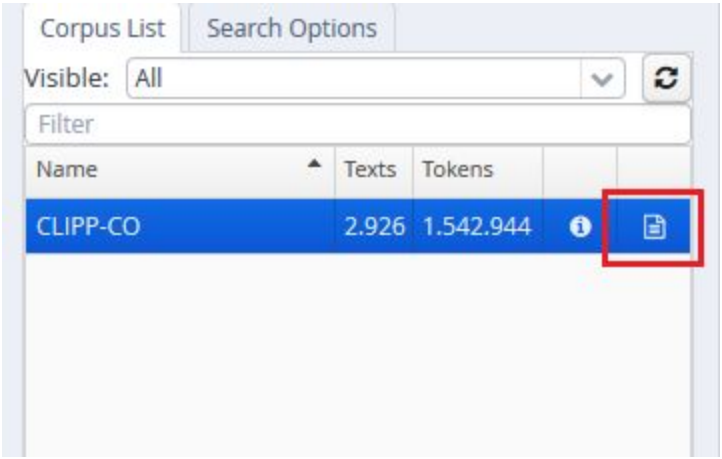
Buscar “Carlos Alvarado” en artículos de opinión de CrHoy:

```
"Carlos" . "Alvarado" & meta::Seccion="Opinion" & meta::Fuente="Crhoy"
```

5. Visualización de texto completo

Es posible visualizar el texto completo de cada artículo, y buscar los artículos por título directamente.

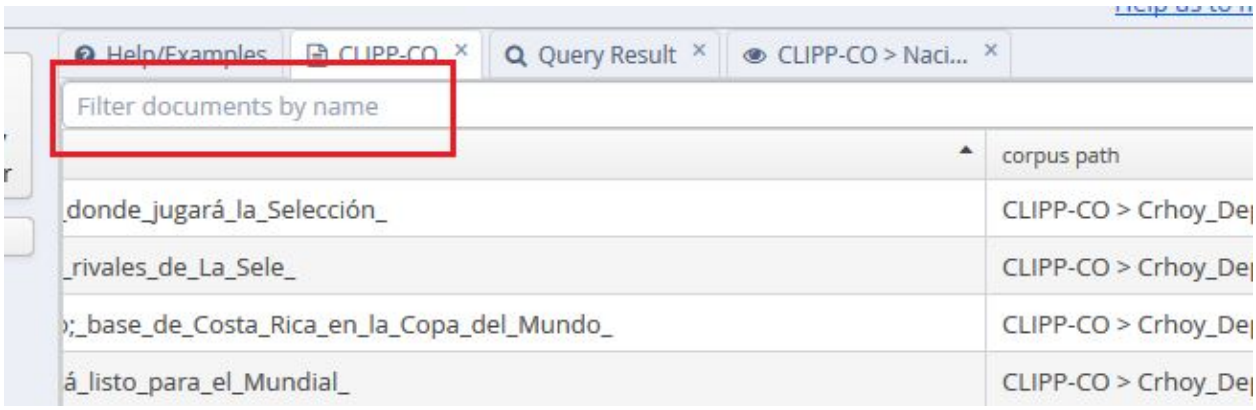
Paso 1:



Paso 2:

NOTA Como los títulos de los artículos son muy largos en este corpus, es probable que tenga que mover la barra inferior hasta que el botón de “full text” aparezca.

Búsquedas:



6. Expresiones regulares básicas

La plataforma de corpus permite el uso de expresiones regulares para realizar búsquedas. Las expresiones regulares se escriben entre / / .

Grupos

Es posible establecer grupos de caracteres con [ ] . Por ejemplo, la siguiente expresión regular:

```
/[Ee]l/
```

Devuelve “El” y “el”. En otras palabras, esta expresión regular busca un token que tenga una letra del grupo [Ee] y luego una “l”.

La siguiente búsqueda:

```
/com[íe]/
```

Devuelve tanto “comí” como “come”.

Operadores * y +

Un asterisco * se utiliza para indicar que el carácter o grupo anterior ocurre 0 o más veces.

Por ejemplo, la siguiente búsqueda indica que “í” puede o no aparecer en la búsqueda:

```
/comí*a/
```

Y devuelve, potencialmente, comía, coma, comíía, comíííía, …

Un + indica que el carácter o grupo aparece una o más veces. La siguiente expresión regular devuelve tanto “pero” como “perro”, y en teoría también “perrro” , y “perrrrrrrrrrrrro”; formas que quizás puedan aparecer en corpora con un registro más coloquial:

```
/per+o/
```

O con grupos:

```
/com[eríaon]+/
```

Devolvería por ejemplo “come”, “comería”, “comí”, “coman”, además de otras posibilidades teóricas como “comoner”, y “comeeeeeee” entre otras infinitas combinaciones.

Comodines

El punto (“.”) actúa como un comodín. Por ejemplo, la siguiente búsqueda regresa, entre otras opciones, “los”, “las”, “les”, etc.

```
/l.s/
```

Los comodines también pueden ser modificados con los operadores * y +.

La siguiente expresión regular:

```
morfo=/.*Fem.*Sing.*/
```

Devuelve, entre otros, el siguiente string de anotación:

ADJ__Gender=Fem|Number=Sing|VerbForm=Part

La siguiente gráfica ilustra qué partes del string son encontradas por qué parte de la epxresión regular:

A	D	J	_	_	G	e	n	d	e	r	=	F	e	m	|	N	u	m	b	e	r	=	S	i	n	g	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	F	e	m	.	.	.	.	.	.	.	.	S	i	n	g	.	.	.