Taller sobre herramientas de análisis textual:

La herramienta Sketch Engine

Facultad de Filología, Universidad Complutense de Madrid

18/02/2019

Autores:

Iván Arias Rodríguez (iarias01@ucm.es)

Ana Fernández Pampillón Cesteros (apampi@filol.ucm.es)

Doaa Samy (doaa_samy@hotmail.com) (doaasamy@cu.edu.eg)

Jorge Arús Hita (jarus@filol.ucm.es)

Objetivo:

El objetivo es proporcionar una introducción al uso de la herramienta *Sketch Engine*. El tiempo de aprendizaje previsto es de 2-3 horas en una única sesión o dos sesiones de hora y media.

Este documento es una actualización de:

Arias Rodríguez, Iván y Samy, Doaa y Fernández-Pampillón Cesteros, Ana María y Arús Hita, Jorge (2017) *Taller sobre herramientas de análisis textual: La herramienta Sketch Engine.* https://eprints.ucm.es/46295/

Sami, D.; Fernández-Pampillón, A.; Arús, J. (2011) "Taller sobre herramientas de análisis textual: la herramienta *Sketch Engine*". Disponible en: http://eprints.ucm.es/13796

Contenido

1. - Introducción

1.1. - Las herramientas de análisis textual: los corpus

Las herramientas de análisis textual permiten el tratamiento automático de la información para apoyar el **estudio empírico de las lenguas**. Permiten la consulta rápida de una o varias **colecciones de textos** electrónicos, así como la preparación y el mantenimiento de bases de datos textuales.

Los textos pueden analizarse desde dos niveles de complejidad:

1. El primer nivel, el **nivel de datos**, es el nivel más simple desde el punto de vista operacional. El objeto de análisis es el **texto puro**, entendido como un conjunto de **caracteres**. Las operaciones automáticas que pueden realizarse son, únicamente, aquellas basadas en la manipulación de los caracteres del texto. Por ejemplo, la localización de todas las palabras que comienzan por un **prefijo determinado**, las palabras que acompañan a otra dada (**colocaciones**) o la **frecuencia** de aparición de un término.
2. En un segundo nivel de complejidad, el **nivel de información** se corresponde con la **interpretación** de los textos. A este nivel corresponden las operaciones que necesitan disponer, además de los textos, de la interpretación de los mismos, como por ejemplo la localización de todos los **verbos transitivos** en una colección de textos o los textos que tratan sobre "la desaparición del atún". Este tipo de **consultas más "inteligentes"** requieren que el **texto** tenga **marcado** de alguna forma este tipo de información. Es importante tener en cuenta que el procesamiento automático de la información es más complicado que el procesamiento de datos y que solo puede realizarse si previamente se ha preparado (**pre-procesamiento**) el texto (datos) insertando la **interpretación (semántica)** de estos datos textuales. Uno de los mecanismos de pre-procesamiento es la inserción en el texto de **marcas o etiquetas** explícitas con la información asociada a cada elemento textual (proceso de etiquetado).

Todas las herramientas de análisis textual permiten el análisis de texto puro (datos), y sólo algunas, como ***Sketch Engine***, ofrecen la posibilidad de analizar también el **texto marcado** (información). Además, *Sketch Engine* dispone de sus propias herramientas para el etiquetado automático del texto, con lo que es capaz de preprocesar el texto puro, convirtiéndolo en texto marcado.

Para trabajar con herramientas de análisis textual hace falta una **colección de textos** representativos de aquello que se desea analizar. A esta colección representativa de textos se le llama un **corpus**. Para un estudio empírico bien fundamentado y unos resultados significativos, es necesario que la colección sea lo suficientemente **grande y equilibrada** como para asegurar que abarca las máximas ocurrencias posibles del fenómeno a estudiar. En este sentido se pueden distinguir varias tipologías de corpus.

1.1.1 - Criterios para recolectar los datos: resumen de tipología de corpus

Se puede establecer una **tipología** de corpus[1] en función de los **criterios** utilizados para la clasificación. Estos criterios pueden ser de diferente índole:

- Según la **modalidad** de la lengua.
  - Corpus de **lengua escrita**.
  - Corpus de **lengua hablada**.
    - **Transcripciones ortográficas** de grabaciones, utilizadas sobre todo en lingüística de corpus.
    - **Grabaciones** y **transcripciones** ortográficas y/o **fonéticas**, empleadas en fonética y tecnologías del habla.
  - Corpus **mixtos**.
- Según la **naturaleza física** de los datos.
  - Corpus **textuales** o escritos: compuestos por una colección de textos.
  - Corpus **orales**: compuestos por una colección de audios.
  - Textos **multimodales** (imagen/video/audio y texto).
- Según la cobertura y la **temática**.
  - Por períodos.
  - Por géneros.
  - Por temas.
  - Por autores.
  - Por registros.
- Según el **número de lenguas**.
  - Corpus **monolingües**: están formados por textos de una sola lengua. Se recopilan con el objetivo de dar cuenta de **una lengua o variedad lingüística**.
  - Corpus **bilingües o multilingües**: están formados por textos de dos (bilingües) o más (multilingües) lenguas **sin que**, en principio, **sean traducciones** unos de otros y sin compartir criterios de selección.
  - Corpus **comparables** ("*paired texts*"): consisten en una selección de textos en **más de una lengua** o variedad lingüística parecidos en cuanto a sus características y que comparten criterios de selección. Se utilizan sobre todo para comparar variedades de la lengua en estudios contrastivos.
  - Corpus **paralelos** ("*bi-texts*"): recogen textos en **más de una lengua** (bilingües o multilingües) pero, a diferencia de los anteriores, se trata del **mismo texto traducido** a una o más lenguas.
- Según la **disponibilidad** y el **modo de recopilación**.
  - Corpus disponibles como recursos en instituciones o en **grupos de investigación**.
  - Corpus construidos a partir de textos seleccionados y **recopilados manualmente** según criterios concretos y para estudiar unos aspectos concretos.

[1] La palabra *corpus* es invariante en cuanto al número (es decir, su plural es igualmente *corpus*). No debe confundirse con la palabra *corpora*, que es el plural de la palabra inglesa *corpus* (que mantiene el plural latino).

- Corpus recopilados automáticamente por herramientas que permiten una búsqueda y recopilación a través de internet mediante el uso de palabras semilla[2].

1.2. - ¿Qué es *Sketch Engine*?

Sketch Engine es una **herramienta de análisis textual en línea** que recibe como entrada un **corpus en cualquier idioma** con, posiblemente, un cierto nivel de **anotación lingüística** para su posterior análisis. *Sketch Engine* ofrece múltiples funciones para el análisis lingüístico:

1) Análisis de colocaciones gramaticales y léxicas (***Word Sketch***): Busca las apariciones en el corpus de ciertas estructuras gramaticales que contienen a una palabra/lema dado, así como coapariciones de términos.
2) Además permite hacer un contraste entre los resultados para las colocaciones de dos palabras (***Diferencia Sketch***), para apreciar las diferencias entre ambas.
3) ***Tesauro***: Crea automáticamente un conjunto de palabras relacionadas con una palabra dada (campo semántico).
4) ***Concordancia***: Permite buscar contextos en los que aparecen ciertas (combinaciones de) palabras (formas flexionadas) y/o lemas.
5) Es posible también realizar una ***concordancia en paralelo*** en corpus multilingües para hallar los equivalentes a palabras y expresiones entre una lengua y otra.
6) ***Lista de palabras***: Realiza conteos de cantidad de apariciones y frecuencias de las palabras/lemas que aparecen en el corpus.
7) Obtención de los ***n-gramas*** (expresiones multipalabra) más característicos del corpus.
8) Extracción de **palabras clave** y términos multipalabra: Se extraen los términos más representativos del texto (en comparación con un corpus de referencia), así como los términos multipalabra más característicos del corpus.
9) ***OneClick Dictionary***: Crea un esbozo de diccionario a partir de un corpus y las relaciones sintácticas de sus palabras.
10) Análisis de ***tendencias*** en el uso de palabras según la fecha de publicación.

Sketch Engine permite trabajar con los siguientes tipos de corpus:

- Corpus que vienen **integrados** en la herramienta.
- Corpus disponibles en **instituciones** o en **grupos de investigación** que hay que "subir" previamente a la herramienta.
- Corpus construidos a partir de **textos seleccionados** y recopilados manualmente.
- Corpus **construidos automáticamente** con la herramienta ***WebBootCat***, que permite la búsqueda y recopilación automática de documentos en Internet.
- Corpus **paralelos**.
- Corpus **etiquetados**.

1.3. - ¿Cómo registrarse y acceder a la cuenta de *Sketch Engine*?

Cualquier persona puede obtener una licencia gratuita temporal con una duración de 30 días. Además, todo aquel que cuente con una dirección de email de la Universidad Complutense puede crearse una cuenta **sin coste alguno hasta marzo de 2022** gracias a la financiación del **proyecto ELEXIS**. Si no se

[2] Una palabra semilla es una palabra representativa de un tema.

posee una dirección de correo de la UCM, el procedimiento para obtener una cuenta gratuita durante los próximos 30 días es el siguiente:

1) Abra la página inicial de *Sketch Engine* en https://www.sketchengine.eu/ y clique en el enlace ***Register*** que aparece en la parte superior izquierda de la página tal y como se muestra en la Figura 1.

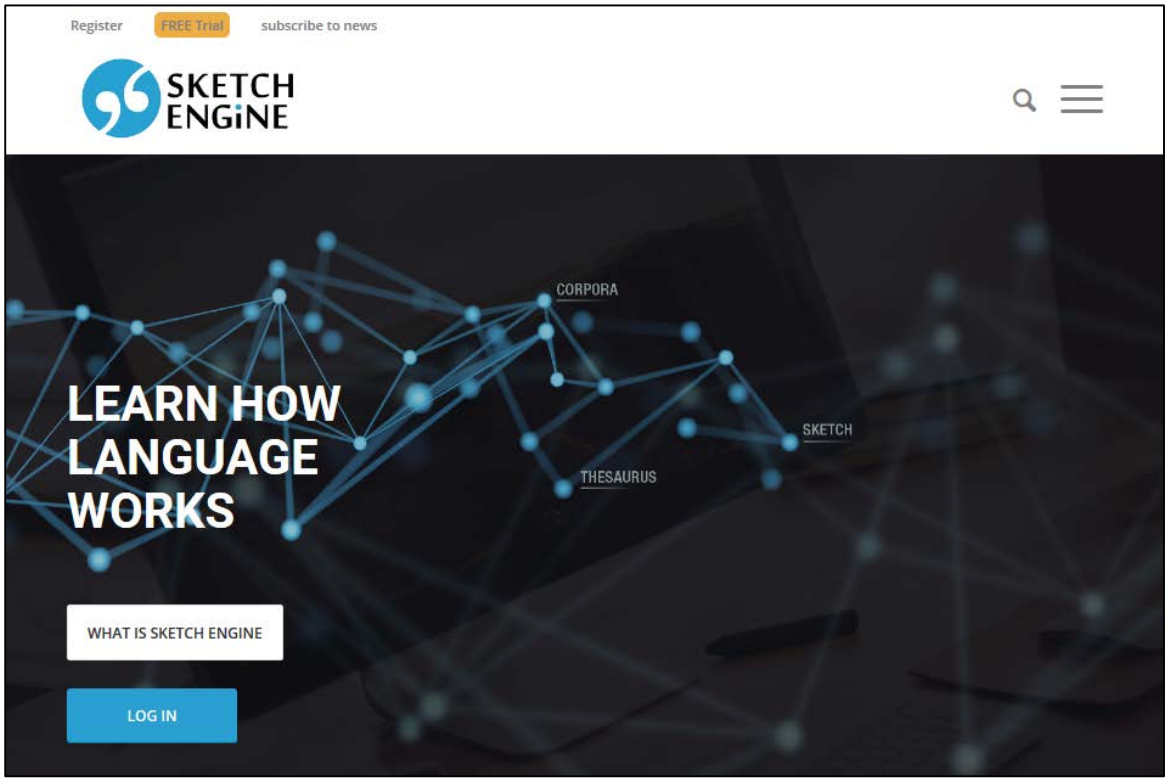

Figura 1: página inicial de Sketch Engine

2) Como aparece en la Figura 2, elija en el menú la opción de ***Free 30-day trial***. Tras ello, rellene el formulario (puede utilizar cualquier dirección de correo) y clique en ***SIGN UP***.
3) Al hacer esto, se le enviará la clave a la cuenta de correo que haya utilizado. Además se le redirige automáticamente a la página de ***Log in***, que se muestra en la Figura 3. Rellene el campo de la cuenta de correo utilizada para darse de alta (o el nombre de usuario que haya escogido) junto con la clave que se le envió y presione el botón de ***LOG IN***. Para futuros usos, puede acceder a la página de *Log in* clicando en el botón azul con el texto *LOG IN* que aparece en la página inicial de *Sketch Engine* (Figura 1).

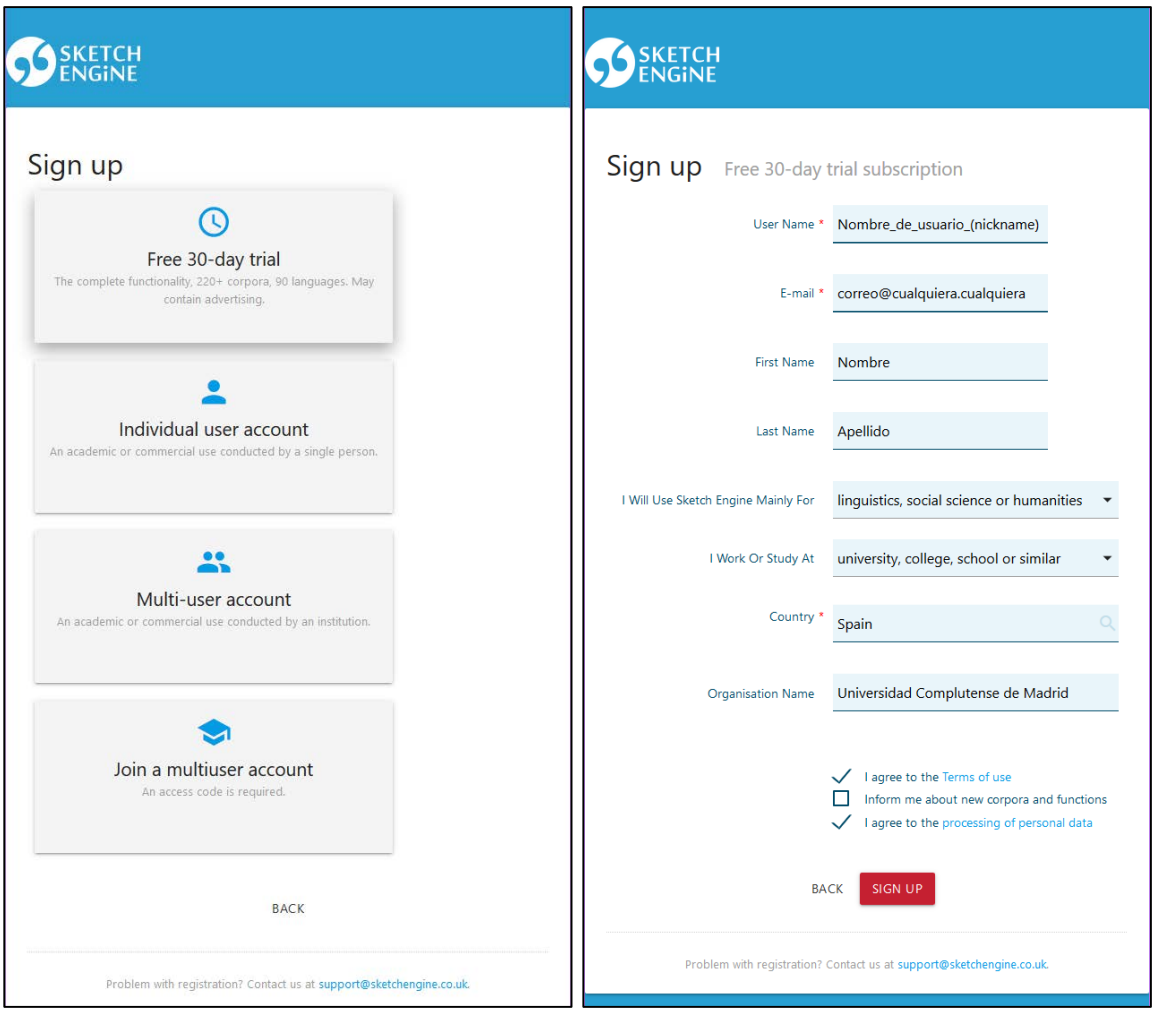

Figura 2: registro con licencia gratuita válida para 30 días

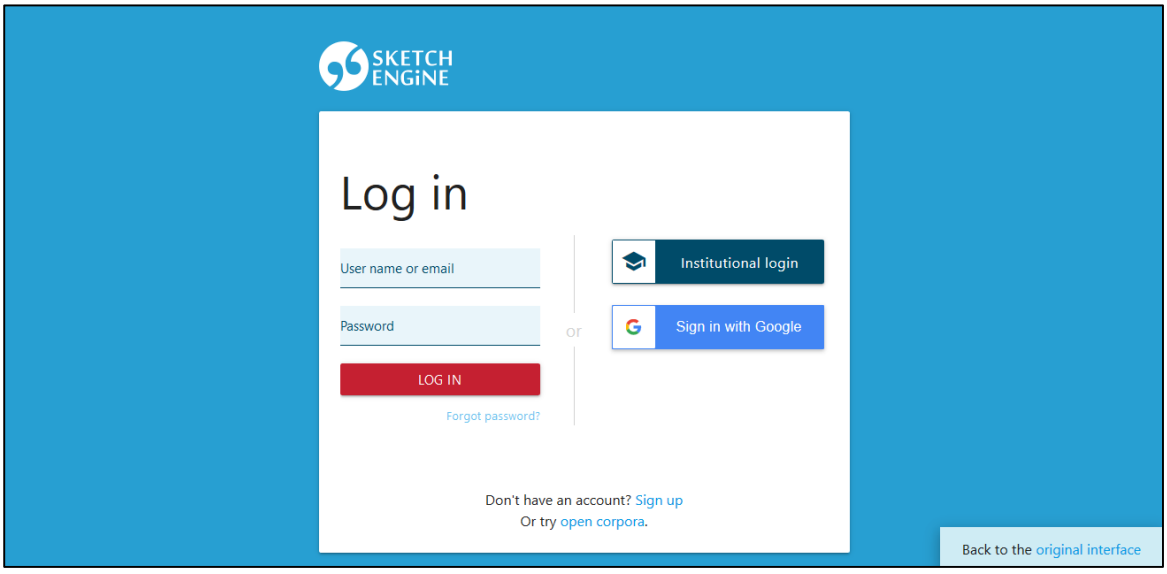

Figura 3: página de Log in

En caso de tener una dirección de correo de la UCM (o de alguna otra universidad incluida en el projecto ELEXIS) se puede registrar como usuario para obtener una **licencia institucional** que será válida hasta marzo de 2022. En este caso, proceda así:

1) Abra la página inicial de *Sketch Engine* en https://www.sketchengine.eu/ y clique en el botón azul de ***LOG IN*** que puede verse en la Figura 1.
2) En la página de *Log in* (Figura 3) clique el botón superior de la derecha que contiene el texto ***Institutional login***.
3) Se le abrirá una ventana emergente como la que se muestra en la Figura 4. Simplemente escriba el texto *Complutense* en el campo de entrada de texto, escoja la institución *UCM – Universidad Complutense de Madrid* y clique el botón (que cambia de *Search* a *Continue* tras seleccionar la UCM). Tras ello, deberá utilizar sus datos de correo de UCM y clave que utilice normalmente y clicar en ***Iniciar sesión***.
4) En la siguiente página se le preguntará si quiere reactivar una cuenta antigua que se haya desactivado o si quiere crear una cuenta nueva. Si nunca ha tenido una cuenta en *Sketch Engine* o si no está seguro pero no le importa que su cuenta se reinicie (si en el pasado hubiera creado su propio corpus, se borraría), clique en ***NEW ACCOUNT***. En caso contrario, clique en *REACTIVATE ACCOUNT* y siga las instrucciones que le aparezcan en pantalla.
5) Seguidamente le aparecerá un formulario como el de la Figura 2 (pero con menos campos) que deberá rellenar de modo similar a como se muestra en dicha figura, para después clicar en ***SIGN UP***.
6) Tras esto, clique en ***GO TO SKETCH ENGINE*** en la ventana emergente para acceder a su cuenta. Además, se le enviará su clave de acceso a su correo institucional de la UCM que deberá utilizar en sucesivos accesos a *Sketch Engine* (clicando el botón azul de *LOG IN* en la página inicial y rellenando sus datos de cuenta de correo y clave de acceso).

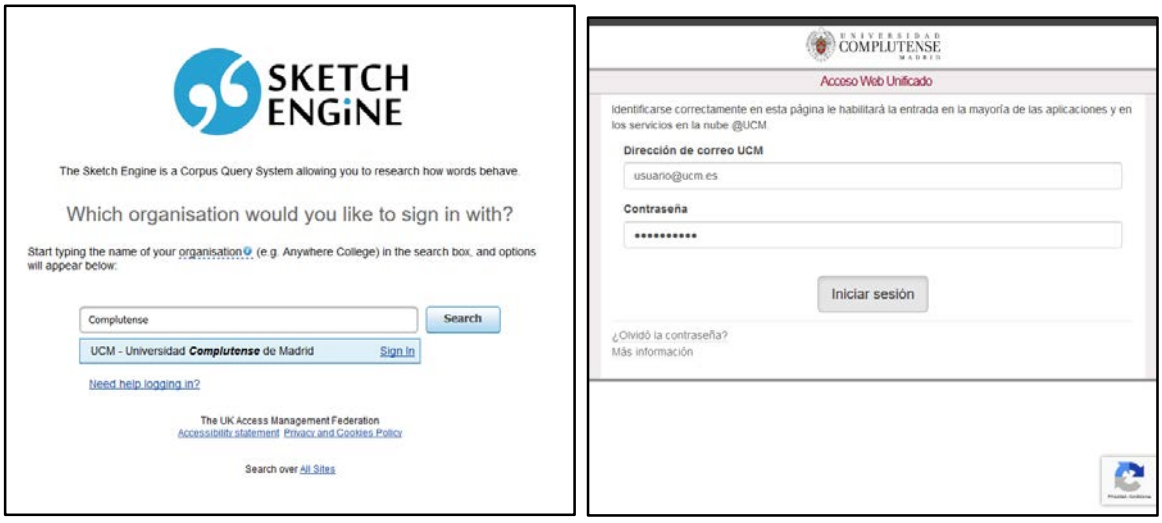

Figura 4: introducción del nombre de la institución y verificación de usuario de la UCM

2. - Selección de un corpus

Para poder empezar a trabajar con *Sketch Engine*, lo primero que tenemos que hacer es seleccionar un corpus sobre el que trabajar. Así pues, la primera tarea es la de seleccionar alguno de los muchos corpus disponibles o crear un corpus propio. Es importante comprender que, una vez **elegido un corpus**, todas las **operaciones** (búsquedas, concordancias…) que se hagan **se realizarán sobre ese corpus**.

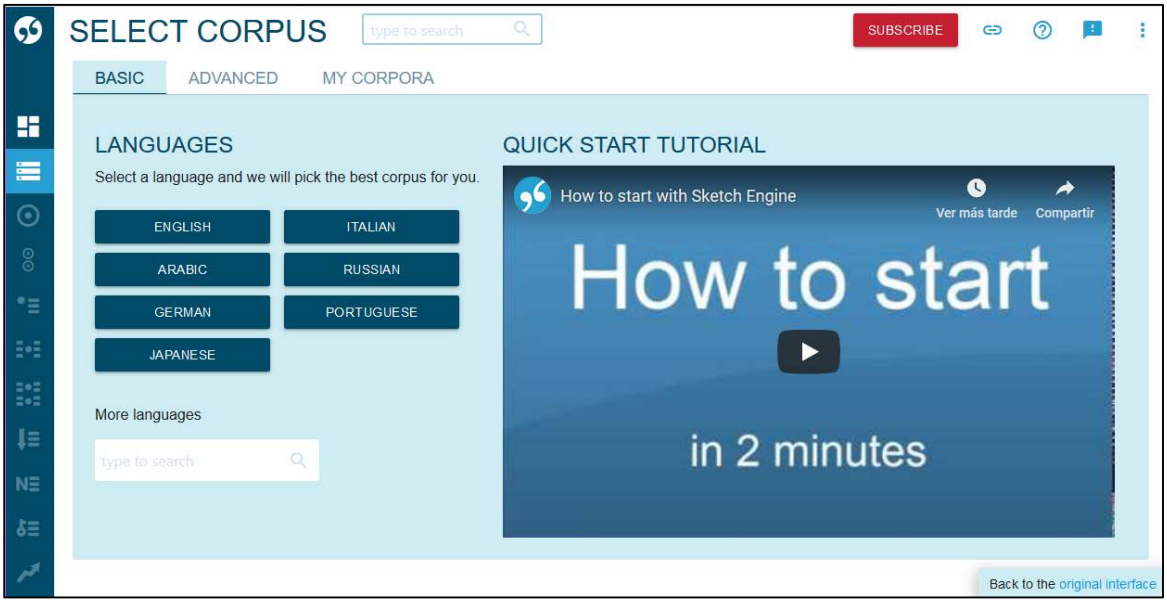

Figura 5: página de selección de corpus de Sketch Engine (Básico)

En el primer acceso a *Sketch Engine*, no habrá ningún corpus seleccionado, con lo que se nos redigirá automáticamente a la **página de selección de corpus**, como se muestra en la Figura 5. Como puede observarse en esta figura, la página muestra once iconos en la parte izquierda de la pantalla, que se corresponden con once herramientas disponibles. No obstante, al no haber escogido ni creado aún ningún corpus, solo las dos primeras opciones están disponibles, mientras las nueve siguientes aparecen atenuadas y no se pueden escoger. La primera de dichas herramientas es el **panel de control** (*Dashboard*). Si clicamos en este primer icono, veremos que no nos permite hacer nada más que seleccionar o crear un corpus, como puede observarse en la Figura 6.

Antes de empezar a usar *Sketch Engine*, hay una serie de acciones que podemos llevar a cabo clicando en los cuatro iconos que aparecen en la parte superior derecha de la pantalla. Es importante hacer notar que el botón rojo de ***SUBSCRIBE*** solo aparece cuando estamos usando una cuenta gratuita válida por 30 días. En caso contrario aparecerá ***Get more space (+)***. De derecha a izquierda, dichos iconos son:

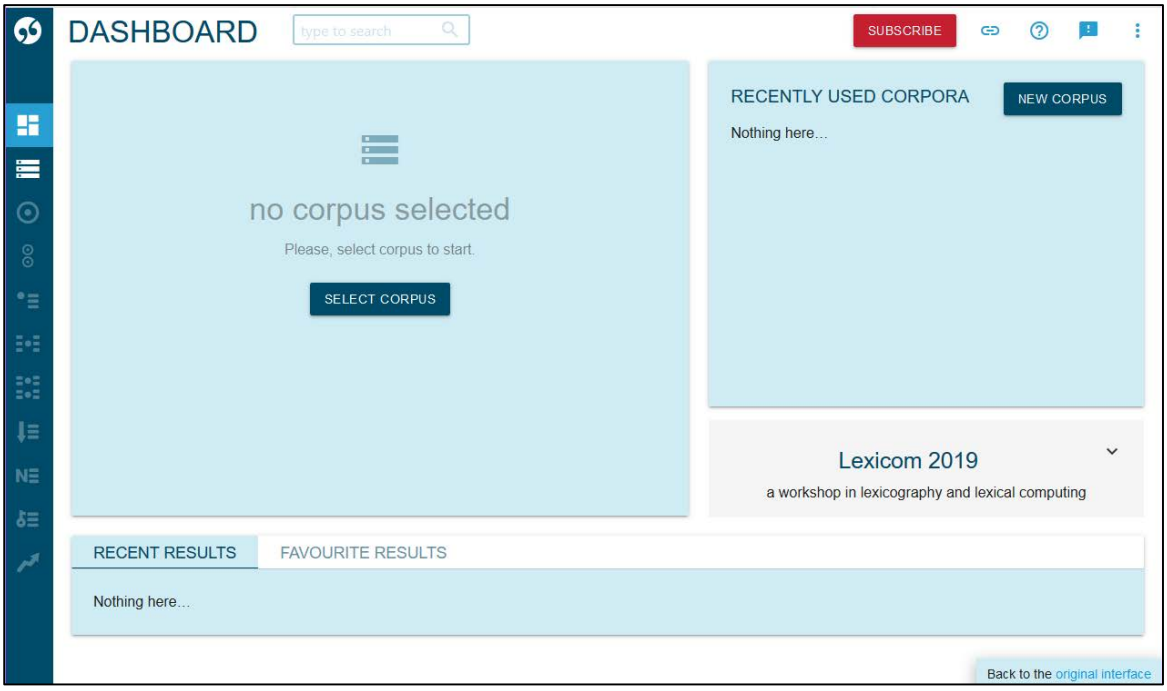

Figura 6: panel de control de Sketch Engine (primer acceso)

- ***Configuración***: Tiene tres opciones.
    a. ***Mi cuenta***: Da información sobre nuestro usuario y el tipo de cuenta que tenemos, así como sobre la cantidad de memoria utilizada en nuestros corpus. Por defecto se permite utilizer hasta 1.000.000 palabras, pero dicho espacio puede aumentarse.
    b. ***Configuración***: Permite cambiar la lengua en la que se muestra *Sketch Engine*, así como ajustar la *densidad* de información (si el texto aparece con más o menos espacio blanco alrededor). En adelante, se utilizará *Sketch Engine* con alta densidad de texto y en español (si bien no toda la herramienta está traducida y seguiremos viendo algunos textos en inglés).
    c. ***Cierre de sesión***: Permite cerrar la sesión abierta en el ordenador. La sesión debe cerrarse al terminar de usar *Sketch Engine* si estamos usándolo en un ordenador compartido.
- ***Comentarios***: Permite hacer consultas a *Sketch Engine* cuando tengamos algún problema con su uso.
- ***Guía de usuario***: *Sketch Engine* dispone de varias guías de usuario, tanto en modo texto como video, que permiten a un recién llegado familiarizarse con el uso de la herramienta.
- ***Enlaces cortos***: Crea enlaces (URLs) cortos a un cierto corpus o página de *Sketch Engine* para enviárselo más comodamente a otro usuario.
- ***Obtener más espacio (+)***: Al clicar aquí se nos da la posibilidad de hacer una petición a *Sketch Engine* para obtener más espacio en nuestra cuenta y así disponer de más de 1.000.000 palabras de almacenamiento para nuestros corpus. Esta opción solo está disponible para usuarios con una cuenta institucional o de pago. Si de verdad nos hace falta el espacio extra de almacenamiento, podemos utilizar esta opción y disponer de más espacio sin coste extra.

Así pues, lo primero que debemos hacer para empezar a utilizar *Sketch Engine* es seleccionar un corpus sobre el que trabajar. Para ello, debemos ir a la página de selección de corpus, en cuya parte superior podemos ver tres pestañas:

Si nos fijamos en la página de selección de corpus, vemos que arriba muestra tres pestañas:

- ***BÁSICO***: Nos permite hacer una selección de un corpus basándose únicamente en la lengua. Si queremos escoger un corpus en una lengua que no sean las siete que aparecen en los botones de la Figura 5, debemos escribir el nombre de la **lengua** (en inglés) en el campo de texto que aparece en la parte inferior izquierda de la pantalla. Si escogemos el español (*Spanish*), se escogerá por defecto el corpus esTenTen[3], un corpus etiquetado que contiene casi 10.000.000.000 palabras (es decir, 10^{10} palabras, de ahí su nombre). Este corpus está etiquetado automáticamente (sin revisión humana posterior) con el etiquetador morfosintáctico FreeLing[4], que tiene una precisión de alrededor del 97% de acierto y que asigna etiquetas basadas en las recomendaciones EAGLES[5].
  Una vez hayamos seleccionado un corpus de trabajo, se nos abren ocho de las diez opciones que hasta ahora estaban desactivadas, tal y como se ve en la Figura 7, que muestra la pantalla de panel de control tras haber seleccionado el corpus español por defecto. Estas mismas herramientas aparecen tanto en el menú de iconos de la izquierda, como en el panel principal del panel de control (donde además se muestra la herramienta para la creación de diccionarios automáticos). También aparece un panel de los corpus recientemente escogidos a la derecha, y un panel de resultados en la parte inferior que se verá más adelante.

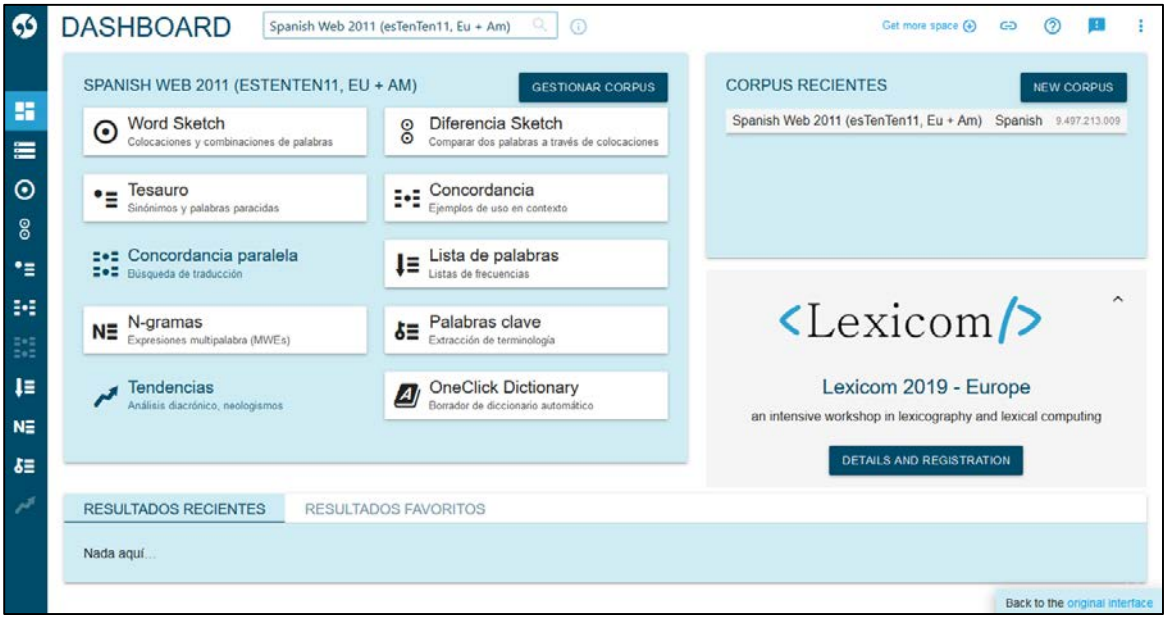

Figura 7: panel de control de Sketch Engine (tras la selección del corpus)

[3] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. y Suchomel, V. (Eds.) 2013. The tenten corpus family.
[4] Padró, L. y Stanilovsky, E. (Eds.) 2012. Freeling 3.0: Towards wider multilinguality.
[5] Leech, G. y Wilson, A. (1996). EAGLES recommendations for the morphosyntactic annotation of corpora.

- ***AVANZADO***: Muestra **todos los corpus** que hay disponibles en *Sketch Engine*. Actualmente existen 492 corpus en 94 lenguas, como puede verse en la Figura 8. La página permite **filtrar** los corpus según la lengua y un gran número de características del contenido del corpus. Los resultados que aparecen en el listado se pueden ordenar en orden creciente o decreciente según la lengua del texto, el nombre del corpus o el número de palabras que contiene.
  El listado, además de mostrar la lengua, el nombre y el tamaño de cada corpus, tiene un botón a la derecha (marcado con el símbolo **…**) que permite obtener ciertos datos y realizar una serie de acciones sobre el corpus en cuestión (salvo la última, las demás solo son aplicables a corpus privados creados tal y como se verá en la siguiente sección):
  a. ***Eliminar corpus***: Borra el corpus y deja de estar disponible en *Sketch Engine*.
  b. ***Editar metadatos de corpus***: Permite modificar los valores sobre la descripción y contenidos del corpus.
  c. ***Aumentar***: Añade contenido a un cierto corpus.
  d. ***Gestionar corpus***: Permite crear subcorpus.
  e. ***Compartir***: Para enviar a alguien un enlace a ciertos corpus.
  f. ***Descargar***: Descarga una copia del corpus en el disco duro local.
  g. ***Ver detalles del corpus***: Da información detallada del contenido del corpus, así como de su etiquetado y de su composición. En la Figura 9 se muestran los detalles del corpus español por defecto esTenTen11.

  Además, en esta pestaña tenemos la posibilidad de crear nuestro propio corpus.
- ***MIS CORPUS***: Se listan los corpus que haya creado el usuario, además de permitir crear nuevos corpus (Figura 10). Mientras no hayamos creado algún corpus propio, esta página permanecerá vacía de contenido.

Figura 8: página de selección de corpus de Sketch Engine (Avanzado)

Figura 9: detalles del corpus español por defecto esTenTen11

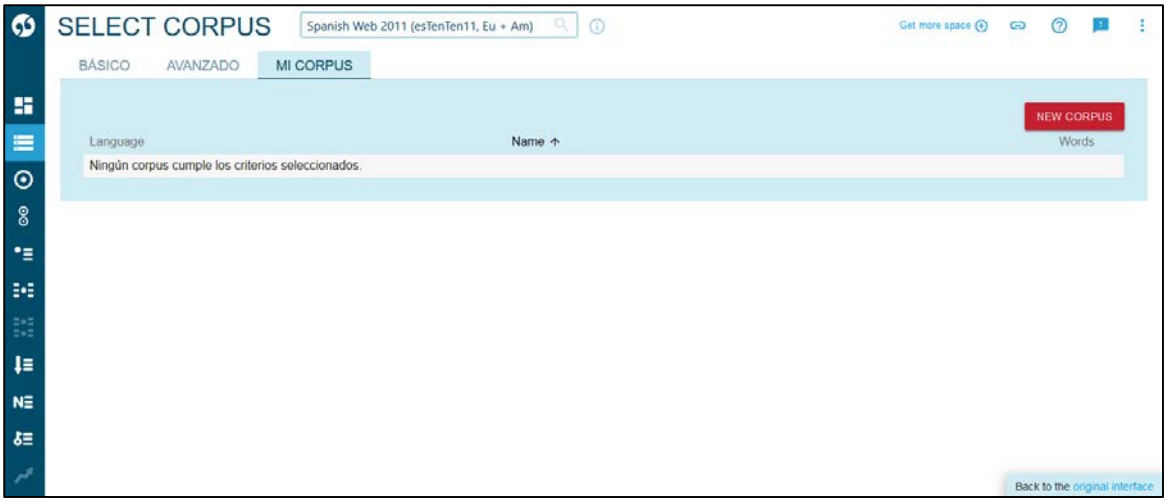

Figura 10: página de selección de corpus de Sketch Engine (Mis corpus)

Cuando queramos cambiar el corpus sobre el que trabajamos, podremos acudir a la página de selección de corpus para utilizar sus filtros. No obstante, en todas las páginas de *Sketch Engine* podemos ver, en la **parte superior**, un campo de texto que muestra el **corpus actualmente seleccionado**, y que permite escoger entre los corpus disponibles filtrando por su nombre. Además, la página del panel de control da la posibilidad de cambiar rápidamente entre los últimos corpus seleccionados, como ya se comentó previamente.

3. - Construcción de un corpus

Para poder utilizar *Sketch Engine*, necesitamos seleccionar uno de los corpus existentes. Además de los casi 500 corpus actualmente disponibles, *Sketch Engine* permite crearnos nuestros propios corpus utilizando distintas estrategias:

- A partir de archivos de texto locales.
- De forma automática, descargando el texto desde Internet:
    - Utilizando una colección de palabras semilla para descargar el corpus desde Internet.
    - Descargando archivos desde un sitio web (indicando simplemente el sitio web, o especificando las páginas concretas).

3.1. - Crear un corpus a partir de archivos de texto locales

Lo primero que debemos hacer es clicar en ***NEW CORPUS***. Este botón está accesible desde varios puntos:

- En el panel de control, un botón azul en la parte superior derecha (Figura 7).
- En la página de selección de corpus, un botón rojo en la parte superior derecha tanto en la pestaña *AVANZADO* (Figura 8) como en la de *MIS CORPUS* (Figura 10).
- A través de la página de gestión de corpus (accesible al clicar el botón *GESTIONAR CORPUS* en el panel de control).

Sketch Engine permite crear corpus monolingües pero también corpus paralelos en varias lenguas. En las dos siguientes subsecciones se explicará cómo crear ambos tipos de corpus.

3.1.1 - Corpus monolingües

Al clicar en *NEW CORPUS*, aparece la pantalla que se muestra en la Figura 11, donde se debe asignar un nombre al corpus que se va a crear, especificar la lengua del texto que se incluirá y añadir una descripción sobre el corpus.

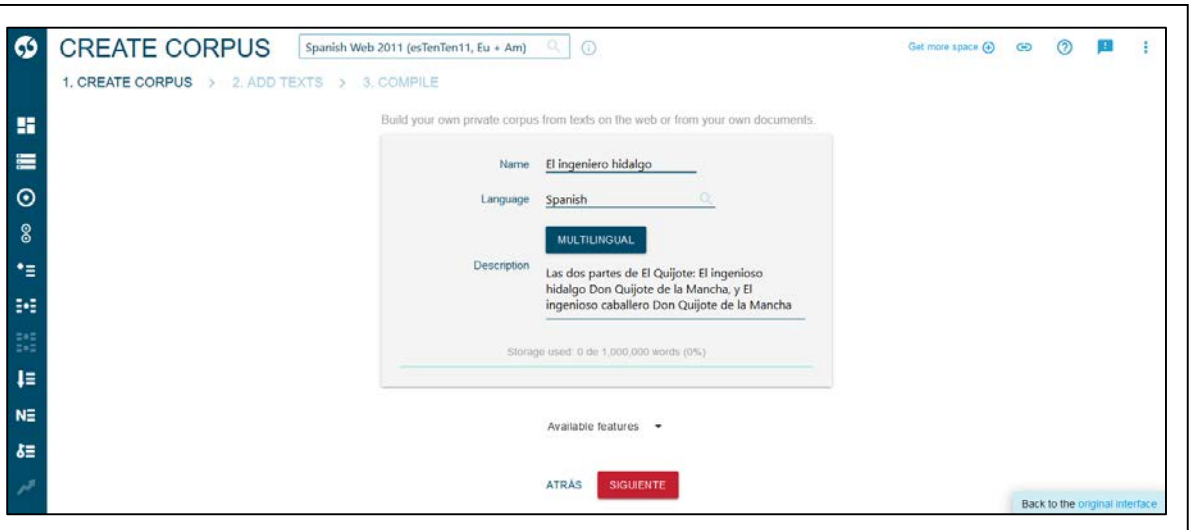

Figura 11: elección del nombre, lengua y descripción del corpus monolingüe que se va a crear

Si clicamos en ***Available features*** se nos muestran las características que tendrá el corpus que vamos a crear. Estas características las produce directamente *Sketch Engine* cuando procese el corpus, y son las que se muestran en la Figura 12.

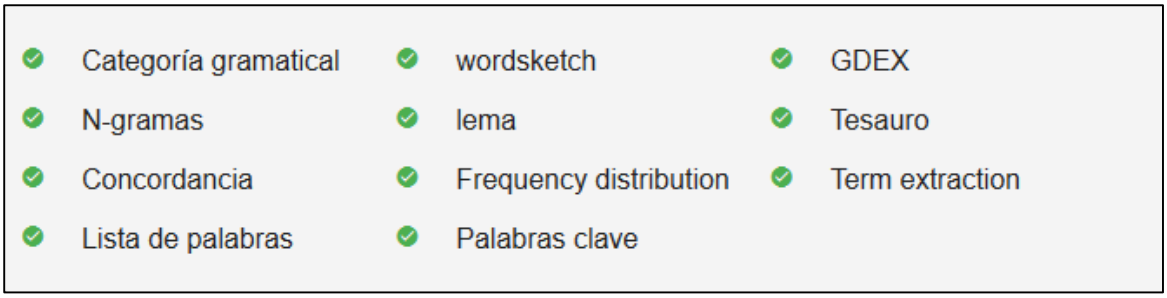

Figura 12: características que tiene un corpus privado creado con Sketch Engine

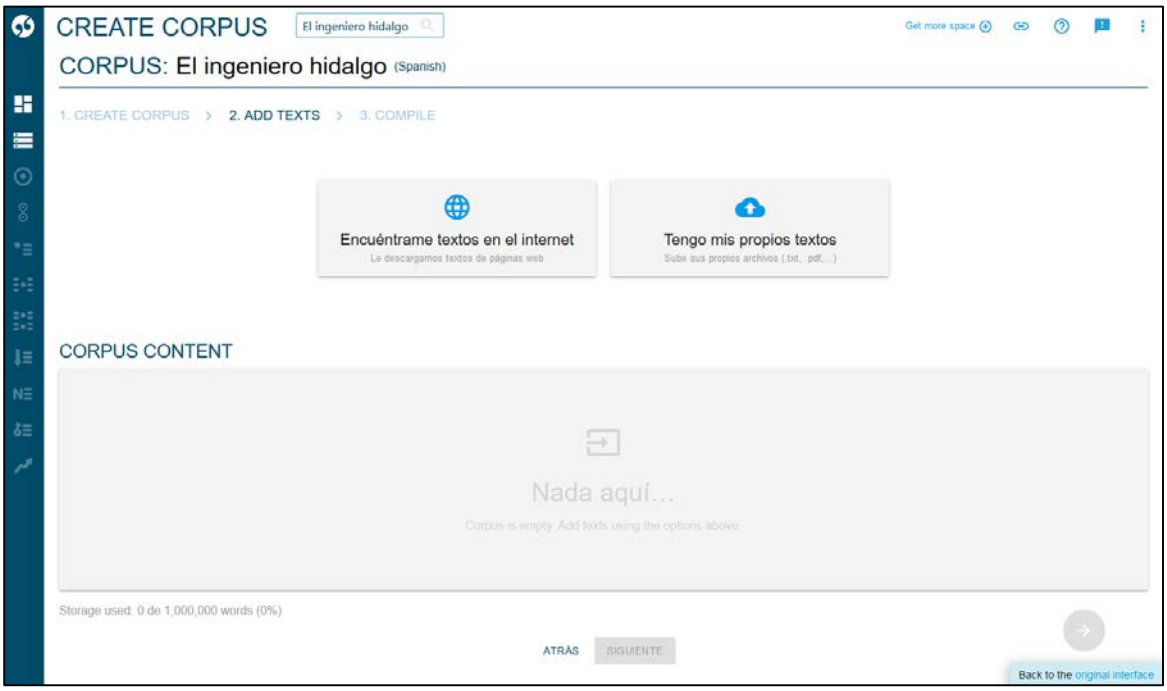

Figura 13: elección del contenido del corpus que se va a crear

Tras ello, se debe clicar en ***SIGUIENTE***, con lo que se llega a la pantalla mostrada en la Figura 13. En este punto, caben dos opciones para elegir el texto que se debe incluir en el corpus:

- ***Encuéntrame textos en el internet***: Esta opción debe utilizarse para crear un corpus automático compuesto por texto descargado de Internet. Su uso se explicará en la siguiente sección.
- ***Tengo mis propios textos***: Se debe escoger esta opción para poder utilizar texto almacenado localmente.

Tras clicar en la segunda opción, se nos pedirá que subamos los archivos que contengan el texto (hasta un máximo de 100). Tras subirlos, *Sketch Engine* procesará el texto, dividiéndolo en *tokens*[6], una tarea que hace a ritmo de unos 5.000.000 de palabras al minuto. Finalmente, la parte inferior se actualizará mostrando la localización y la cantidad de *tokens* que tiene el corpus hasta el momento (Figura 14).

Figura 14: estado actual del corpus que se está creando

Sketch Engine puede manejar muchos tipos distintos de archivos: .doc, .docx, .htm, .html, .ods, .pdf, .tei, .tmx, .txt, .vert, .xlf, .xliff, .xls, .xlsx y .xml. Además, es posible también seleccionar **archivos comprimidos** de tipo .zip, .tar.bz2, .tar.gz o .tgz que contengan más de un archivo en su interior (todos aquellos archivos comprimidos que no tengan alguna de las extensiones descritas anteriormente, se ignorarán).

En vez de subir un archivo, también es posible copiar y pegar un texto directamente. Para ello basta con clicar en el enlace ***or paste text*** que aparece en la parte superior derecha de la pantalla. Al clicar se nos abre un cuadro de texto donde podremos pegar el texto deseado.

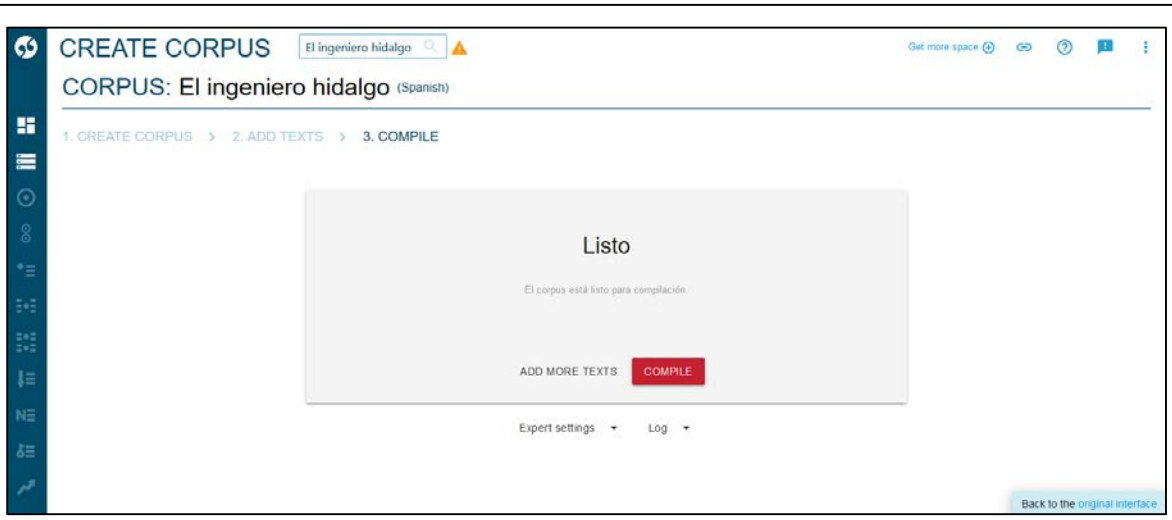

Figura 15: corpus privado listo para compilar

Cuando se haya subido todo el texto, se debe clicar en el botón rojo ***SIGUIENTE*** situado en la parte inferior de la pantalla lo que nos lleva a la pantalla mostrada en la Figura 15. En caso de que nos

[6] *Token* es un término utilizado en Lingüística Computacional y Procesamiento del Lenguaje Natural para referirse a un elemento de texto (habitualmente separadas por espacios) que ha de procesarse de manera independiente. Un *token* puede ser una palabra, pero también lo son los signos de puntuación, números, símbolos, etc.

hayamos olvidado de incluir algo, se puede clicar en ***ADD MORE TEXT*** para volver al punto mostrado en la Figura 13. También se pueden modificar varios ajustes que afectan a la compilación del corpus si clicamos en ***Expert settings***, lo cual despliega un menú que se muestra en la Figura 16. En él podemos elegir si se elimina el contenido duplicado que pueda haber en el corpus, el tipo de etiquetas morfosintácticas que se usarán (bien tipo FreeLing, basadas en las recomendaciones EAGLES, bien las Universal Tags, o bien ninguna), y otros valores. Es recomendable utilizar los valores que aparecen por defecto. Por ejemplo, en caso de no etiquetarse, muchas de las herramientas de *Sketch Engine* no podrán utilizarse.

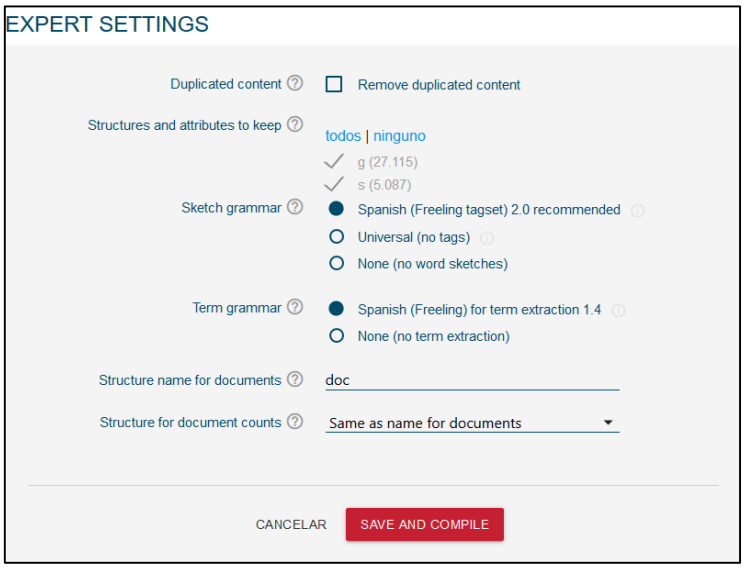

Figura 16: ajustes previos a la compilación de un corpus privado

Tras haber añadido todo el texto al corpus, este se debe compilar clicando en el botón ***COMPILE***. El proceso de compilación es el que aporta la información extra al corpus recién creado, principalmente gracias al etiquetado morfosintáctico, al lematizado y al análisis sintáctico, gracias a los cuales se realiza también un análisis acerca de patrones de aparición de lemas y su posición y función en la oración. Es necesario realizar la compilación antes de poder utilizar este corpus, aunque no es necesario hacerlo inmediatamente después de haber creado el corpus: su compilación puede quedar como pendiente (en caso de que aún no esté creado completamente) y realizarse más adelante.

El proceso de compilado puede llevar segundos o minutos, según sea su tamaño. Pasado este tiempo, el corpus ya estará **disponible** para su uso, con lo que aparecerá como un corpus más, tanto en el listado de corpus propios del usuario, ***MIS CORPUS***, o en el listado de todos los corpus, ***AVANZADO*** (mostrados respectivamente en la Figura 10 y la Figura 8).

3.1.2 - Corpus multilingües

Para crear un corpus multilingüe, deberemos clicar en el botón ***MULTILINGUAL*** que puede verse en la Figura 11, tras lo cual se muestra el menú que aparece en la Figura 18. Lo único que hace falta es ponerle un nombre al corpus y adjuntar un archivo con el texto paralelo. Dicho archivo puede ser de formato .txt .tmx, .xls, .xlsx o .xliff y debe contener en cada línea el texto alineado (pueden ser frases o párrafos) en las dos o más lenguas que contenga. La primera línea debe contener el nombre en inglés de la lengua en cuestión. Tras escoger el archivo se debe pulsar en el botón ***SIGUIENTE*** y se nos

mostrará el menú que aparece en la Figura 18. Al crear un corpus multilingüe se crearán tantos corpus monolingües como lenguas haya en el archivo que hayamos utilizado, con lo que cada uno tendrá su propio nombre y tendrá especificada su propia lengua.

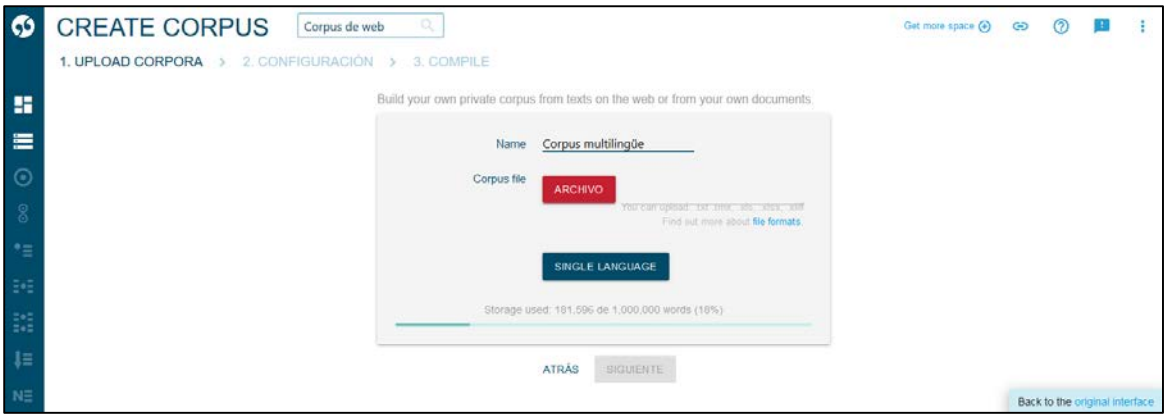

Figura 17: ajustes previos a la compilación de un corpus multilingüe

Figura 18: selección de idiomas del corpus multilingüe

El último paso consiste en pulsar el botón ***SIGUIENTE*** para que *Sketch Engine* compile los distintos corpus monolingües que componen el corpus multilingüe, y estén disponibles en la lista de corpus.

3.2. - Crear un corpus automático a partir de texto descargado de internet

No es necesario tener localmente el contenido del corpus que se vaya a crear, sino que dicho contenido se puede extraer directamente de Internet. Si queremos añadir este tipo de contenido a un corpus, deberemos proceder tal y como se explicó en la sección anterior, hasta llegar al punto mostrado en la Figura 13, momento en el cual deberemos clicar en ***Encuéntrame textos en el internet***. Es importante notar que no hay ningún problema en crear un corpus que contenga texto extraído de internet junto con texto subido de archivos locales o el portapapeles.

Tras clicar en dicha primera opción se nos muestra el menú de la Figura 19, donde se debe elegir una de las tres maneras en las que se añadirá texto al corpus automáticamente:

1) Utilizando **palabras semilla**.
2) Descargando el texto de una **lista de direcciones web** (URLs) que indican páginas concretas.

3) Eligiendo **una web concreta** y dejando que *Sketch Engine* sea quien encuentre y descargue las páginas que se deben utilizar.

Estas tres formas se explican en las siguientes subsecciones.

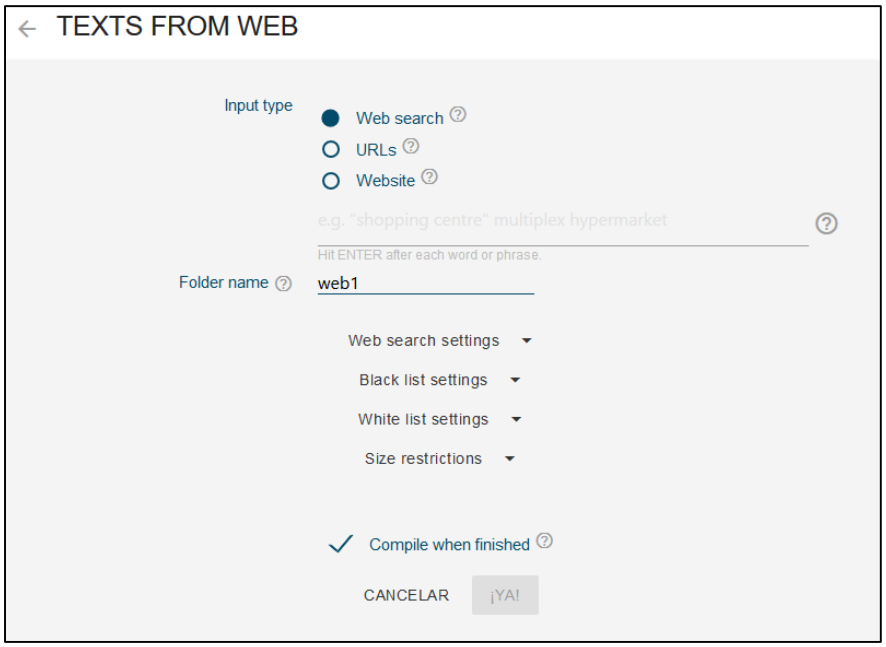

Figura 19: menú para añadir texto a un corpus extraído directamente de Internet (palabras semilla)

3.2.1 - Uso de palabras semilla

Si se escoge el primero de los tres métodos en la Figura 19 (***Web search***), se deberán incluir en el campo de texto las palabras semilla, separadas por comas, que determinarán los archivos que se descargarán de forma automática. Se aconseja utilizar entre 3 y 20: cuantas más palabras semilla se definan, mayor será el número de páginas web descargadas y más amplio será el corpus (será mayor en tamaño y contendrá texto más diverso), ya que se utilizarán **subgrupos de tres** de dichas palabras semilla y con ellas se buscarán contenidos en internet usando el motor de búsqueda Bing.

Hay una serie de ajustes en relación a la manera en la que se extraerá el corpus:

- ***Web search settings***: Se puede escoger **cuántas páginas** se descargan en cada búsqueda realizada, además de la posibilidad de incluir una lista de **sitios web** que serán los únicos de los que se descargarán páginas.
- ***Black list settings***: Evita descargar páginas que contengan más palabras semilla que un cierto límite, y también aquellas que tengan más de un cierto número de palabras semilla distintas. Además, se puede introducir una lista de términos de forma que no se descargará ninguna página que contenga alguno de dichos términos.
- ***White list settings***: Al contrario que en el grupo anterior, aquí se especifica el número mínimo de palabras semilla (distintas o no) que debe tener una página para aceptarse, y puede incluirse una lista de términos que tienen que estar incluidos en la página para ser aceptada. También se puede indicar cuál debe ser el ratio mínimo entre el número de palabras totales en el documento y el número de palabras clave.

- ***Size restrictions***: Se pueden imponer límites de tamaño superior e inferior de los documentos descargados antes o después de limpiarlos de todo contenido no estrictamente textual.

Además, hay una opción, activada por defecto, *Compile when finished* que, como indica, procederá a la **compilación automática del corpus** tras la descarga y preprocesado del texto.

Cuando estén fijadas las palabras clave y todos los ajustes, se debe clicar el botón ***¡YA!*** y esperar a que *Sketch Engine* descargue y procese las páginas, y cree finalmente un corpus con todo ese texto. Según la cantidad de palabras semilla que se haya utilizado, este proceso puede durar un tiempo considerable (desde algunos minutos, hasta horas) durante el cual se va mostrando el progreso. Al acabar, el corpus queda disponible para su uso.

No es necesario esperar a que el corpus se haya descargado para poder hacer otras cosas a la vez: la descarga se realiza desde el servidor de *Sketch Engine* y no desde el propio ordenador, con lo que se puede cerrar la ventana o incluso hacer otras tareas en *Sketch Engine* mientras el corpus crece.

3.2.2 - Uso de URLs concretas

Este método es el **menos automático** de los utilizados para crear un corpus descargando texto de Internet: crea el corpus a partir de una serie de documentos cuyas direcciones web (URLs) se indican de forma explícita. Basta utilizar el segundo de los tres métodos que aparecen en la Figura 19 (*URLs*), con lo que se mostrará el menú que aparece en la Figura 20.

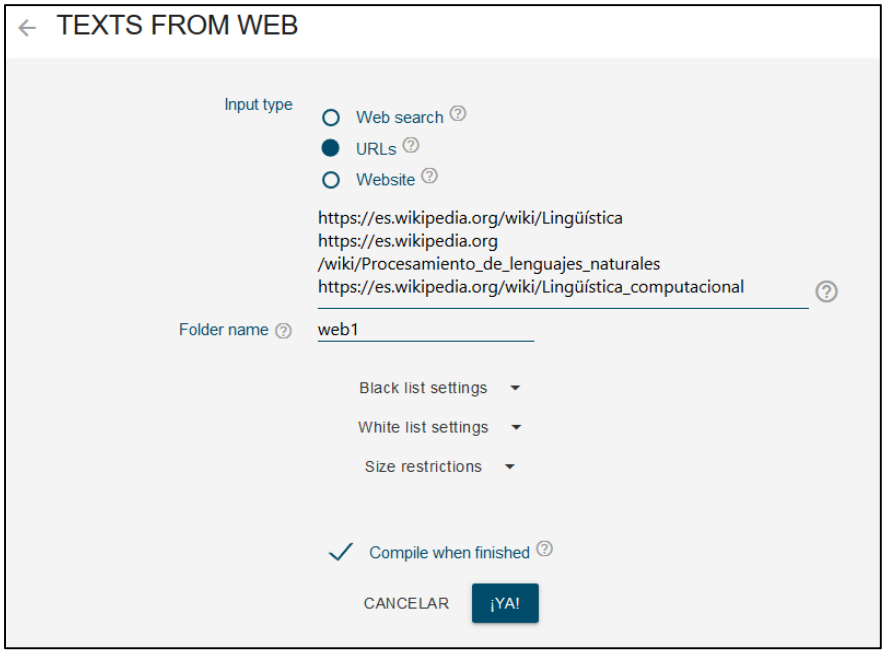

Figura 20: menú para añadir texto a un corpus extraído directamente de Internet (URLs)

Las direcciones web deben escribirse en el campo de texto, separadas por espacios. Y se tienen unas opciones similares a las explicadas en la subsección anterior. Tras rellenar todos los datos, basta con clicar el botón ***¡YA!*** y esperar a que se descarguen y procesen las páginas webs indicadas para que el corpus aparezca en el listado y pueda comenzar a utilizarse.

3.2.3 - Uso de un *Website*

Otra posibilidad a la hora de crear un corpus automático, es la de **elegir un *Website*** y dejar que *Sketch Engine* **encuentre páginas alojadas en dicha web**, y las utilice para crear un corpus. Esto es interesante, por ejemplo, para crear un corpus con artículos de un periódico online.

Para ello, se escoge el tercero de los tres métodos en la Figura 19 (*Website*). Al hacerlo, se muestra un campo de texto (como se ve en la Figura 21), en el que se debe introducir la **dirección de la web** (por ejemplo, www.elpais.com). También es posible indicar escoger un *path* concreto dentro de una página web, como www.elpais.com/internacional/, de forma que *Sketch Engine* solo buscaría en este caso aquellas páginas de la web www.elpais.com que estén **bajo el subdirectorio** /internacional/ (es decir, solo páginas de noticias internacionales). Por lo demás, las opciones que se dan son las mismas que en la subsección anterior.

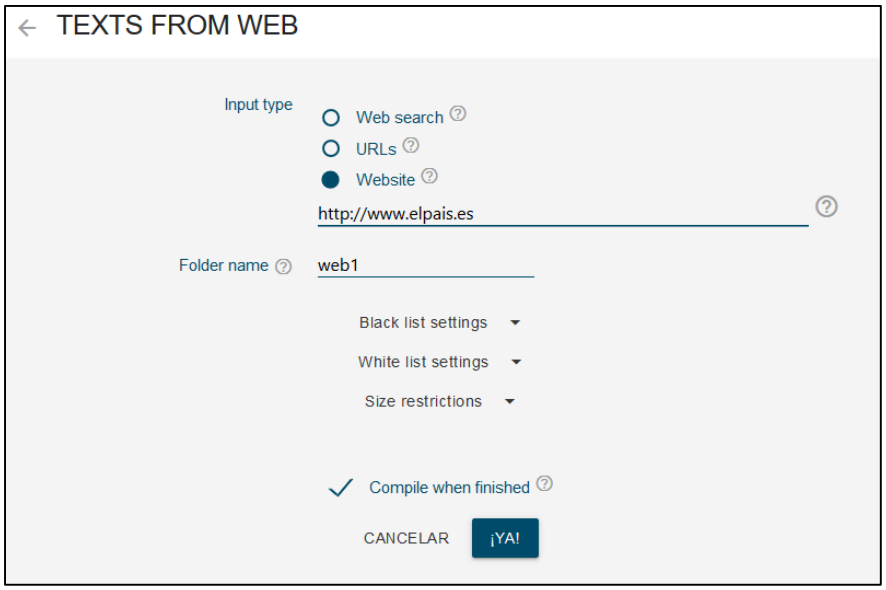

Figura 21: menú para añadir texto a un corpus extraído directamente de Internet (Website)

Tras especificar la página web de la que se extraerá el texto y seleccionar los ajustes, se clica en *¡YA!*, y como en los casos anteriores, se debe esperar hasta que haya terminado el proceso. *Sketch Engine* procurará buscar en la web indicada y extraer de ahí hasta **2.000 páginas**, con lo que el proceso puede alargarse durante **horas**.

Es importante tener en cuenta que usualmente las páginas web no desean gastar tráfico de datos sirviendo páginas a procesos robot como el que utiliza *Sketch Engine*, que se dedica a descargar de forma automática el contenido de una web. Es por ello que el uso de este método **no funcione correctamente en muchas páginas**.

3.3. - División en subcorpus

Muchas veces es interesante obtener datos de una cierta parte de un corpus y no de su totalidad. Para hacer esto se puede dividir un cierto corpus en varios subcorpus, de forma que tengamos acceso únicamente a ciertos contenidos a la hora de analizarlo. Para ello, teniendo seleccionado el corpus que queremos subdividir, clicamos en el botón de ***GESTIONAR CORPUS*** que aparece en la parte superior

central de la página del panel de control (Figura 7). Tras hacer esto, se nos muestra la página de gestión de corpus que se ve en la Figura 22, donde deberemos clicar en ***Subcorpora***.

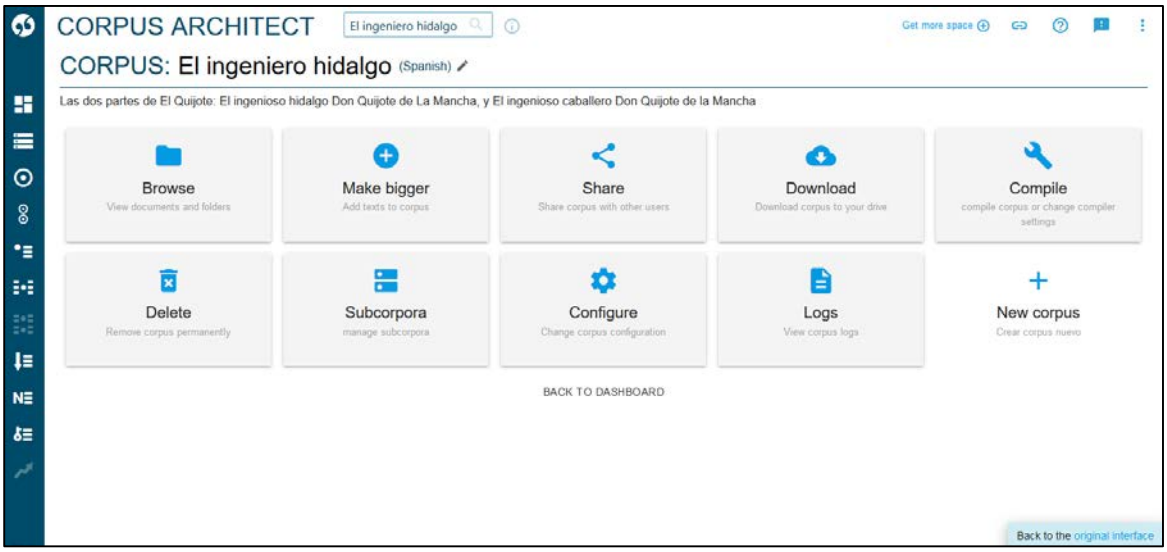

Figura 22: página de gestión de corpus

Si el corpus escogido no había sido subdividido previamente, no contendrá ningún subcorpus, y se nos mostrará una pantalla como la que aparece en la Figura 23. Para crear un subcorpus debemos clicar en el botón ***CREATE SUBCORPUS***, lo que hace aparecer el menú mostrado en la Figura 24, donde se ha clicado en ***Expandir todo*** para mostrar el contenido el corpus. Basta con escribir el nombre del subcorpus y seleccionar los documentos que queremos que contenga (clicando sus nombres o IDs).

Figura 23: subcorpus en el corpus seleccionado (vacío)

Figura 24: menú para escoger el contenido del subcorpus

Se pueden crear tantos subcorpus como se desee, que pueden tener contenido solapado entre ellos. Cuando lo hayamos hecho, sus nombres y tamaños aparecen en la página de subcorpus (Figura 25) que antes se mostraba vacía.

Figura 25: subcorpus en el corpus seleccionado (con dos subcorpus)

El tener un corpus dividido en varios subcorpus nos capacita para que *Sketch Engine* realice una serie de funciones únicamente sobre estas partes del corpus y no sobre el total.

4. - Explorar el corpus

La página de selección de corpus de *Sketch Engine* (mostrada en la Figura 8) muestra los corpus de los que se dispone. La forma de trabajar en *Sketch Engine* es la de **escoger el corpus** sobre el que se va a trabajar, y una vez elegido, **todas las operaciones se hacen sobre ese corpus** (hasta que se escoja un corpus distinto). En todo momento se nos muestra el corpus escogido en la parte superior de todas las páginas de *Sketch Engine*.

Sobre este corpus seleccionado, *Sketch Engine* nos muestra en la pantalla de panel de control (Figura 7) una serie de herramientas disponibles. De ellas, hay ocho que están disponibles para todos los corpus monolingües que hayan sido compilados con *Sketch Engine* y que son las siguientes:

1) ***Word Sketch***: Busca **patrones sintácticos** estadísticamente relevantes de un lema.
2) ***Diferencia Sketch***: **Compara los patrones sintácticos** de dos lemas.
3) ***Tesauro***: Construye el **tesauro** o vocabulario de palabras relacionadas semánticamente con un cierto lema dado.
4) ***Concordancia***: Busca **concordancias** –las palabras deseadas, junto con todas las citas de los lugares en las que se hallan– simples o avanzadas.
5) ***Lista de palabras***: Obtiene la lista de palabras con sus **frecuencias** de aparición en el corpus.
6) ***N-gramas***: Extrae las expresiones multipalabra más típicas del corpus.
7) ***Palabras clave***: Obtiene la lista de **palabras clave** (y términos multipalabra) del corpus.
8) ***OneClick Dictionary***: Crea un esbozo de diccionario a partir de un corpus.

Además, hay otras dos herramientas que necesitan tener seleccionado un corpus con unas ciertas características especiales:

9) ***Concordancia paralela***: El corpus seleccionado debe ser un corpus paralelo en dos o más lenguas. Establece correspondencias entre palabras de dos lenguas.
10) ***Tendencias***: Se necesita un corpus cuyos documentos estén etiquetados con la fecha de publicación. Permite obtener datos sobre la variación en el uso de ciertas palabras en la lengua del corpus.

Estas herramientas se explicarán a lo largo de las siguientes secciones.

4.1. - Word Sketch: Relaciones gramaticales

Esta opción, permite explorar los patrones sintácticos de un lema concreto proporcionando **información sobre la posición y función sintáctica** en que aparece el lema en cuestión. Al clicar el botón *Word Sketch* se muestra un menú con dos pestañas en la parte superior: BÁSICO y AVANZADO. Nos centraremos en esta segunda pestaña, que es la que se muestra en la Figura 26. Ahí se debe rellenar el campo *lema*, que puede rellenarse con uno o más lemas. También se tienen otras opciones para escoger, como la **categoría gramatical** que se busca (puede haber lemas coincidentes para más de una categoría gramatical, como “ante” o “bajo”, que pueden ser preposiciones o un nombre/adjetivo), la frecuencia mínima de aparición o la puntuación mínima. La puntuación es el resultado de una operación matemática en la cual se mide cuántas veces aparece el lema en el corpus

teniendo una cierta relación gramatical con otra palabra, en relación a cuántas veces aparecen por separado. También se puede restringir la búsqueda a un cierto **subcorpus** del corpus seleccionado.

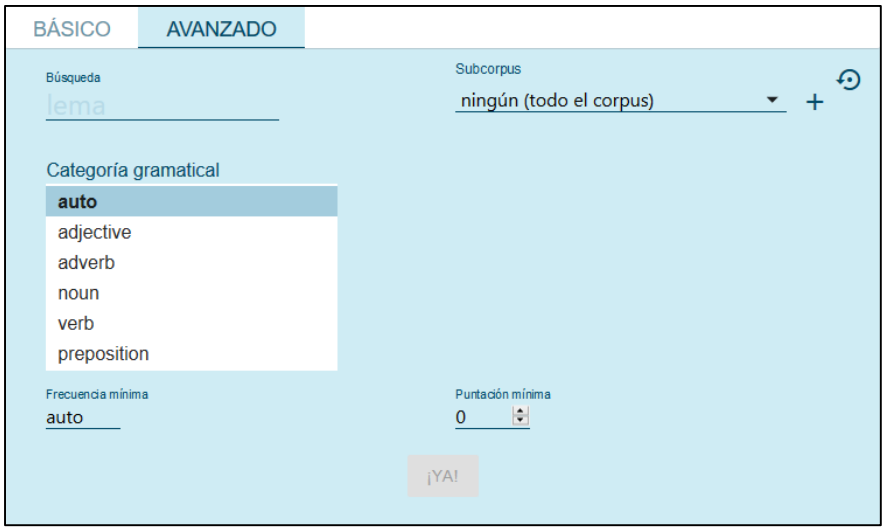

Figura 26: menú para Word Sketch (AVANZADO)

El resultado del *Word Sketch* para "caballero" (en un corpus que está formado por el texto de las dos partes de El Quijote), aparece en la Figura 27. En la imagen se puede observar como "andante" es el **modificador más usual** de "caballero". Se puede observar también muchas otras columnas[7], en las que se muestran los verbos que tienen a "caballero" **más frecuentemente como objeto** (en este caso, "armar"), ciertas otras relaciones sintácticas, y las **combinaciones más usuales con preposiciones** y verbos copulativos, por ejemplo. Cada columna representa una relación gramatical.

El menú desplegable de la parte superior izquierda nos indica que los resultados que se muestran son los que se obtienen cuando el lema actúa como una cierta categoría gramatical, y cuántas apariciones tiene. Si clicamos en él podemos observar que en el corpus aparece "caballero" como adjetivo 162 veces. No obstante, ya que el etiquetado es automático, podemos comprobar cómo muchos de dichos resultados muestran al lema "caballero" actuando como nombre y no como adjetivo.

A la derecha de dicho menú desplegable se tiene un botón marcado con el icono **…** que nos permite pasar directamente a las herramientas de Corcondancia y de Tesauro, que se explican más adelante en sus respectivas secciones. En la parte superior derecha se tienen seis iconos, cuya utilidad es la siguiente (de izquierda a derecha):

[7] Debido a las limitaciones en cuanto al número mínimo de ocurrencias, que se fija en el menú de la Figura 26, aparecen relativamente pocas columnas. En corpus mayores se obtienen resultados mucho más detallados.

Figura 27: resultado de Word Sketch para "caballero"

- ***Cambiar criterios***: Permite modificar los criterios de búsqueda, haciendo aparecer el menú que se vio en la Figura 26.

- ***Descargar***: Descarga los resultados que se están visualizando como un archivo de extensión .csv, .xls, .xlm o .pdf.
- ***Cambiar ajustes de visualización***: Permite mostrar o no, para cada resultado que se muestra en las tablas, la frecuencia de aparición (en realidad, el número de apariciones), la puntuación y el ejemplo más común en el corpus. Además, permite ordenar los resultados de cada tabla según la puntuación o según la frecuencia, así como agrupar resultados semánticamente similares. Al habilitar esta opción se nos mostrará una barra de desplazamiento en la que podremos escoger cómo de parecidos deben ser los lemas para agruparse.
- ***Mostrar visualización***: Al clicar en este icono se nos muestra una interesante visualización gráfica de los resultados de las tablas. El gráfico es interactivo y tiene además varios ajustes de configuración a la derecha que permiten añadir o eliminar características del gráfico. Para el resultado obtenido de "caballero", dicha representación gráfica se muestra en la Figura 28.
- ***Detalles de pantalla***: Muestra los criterios de búsqueda que se están utilizando.
- ***Agregar a favoritos***: Si clicamos en este icono, se guardará esta búsqueda en nuestros favoritos, y se hará accesible desde la pestaña ***RESULTADOS FAVORITOS*** que aparece en la parte inferior del panel de control (Figura 7).

Además, cada una de las tablas que se muestran como el resultado de aplicación de la herramienta *Word Sketch* tiene en su parte superior cuatro iconos:

- El que está más a la izquierda, permite cambiar a nuestro gusto la **posición** en la que aparecen las distintas tablas.
- El segundo icono permite ver las **colocaciones** de los resultados mostrados en dicha tabla.
- El tercer icono **elimina todas las demás tablas** de la visualización y deja visible solo esa tabla.
- El que está más a la derecha **elimina dicha tabla** de la visualización.

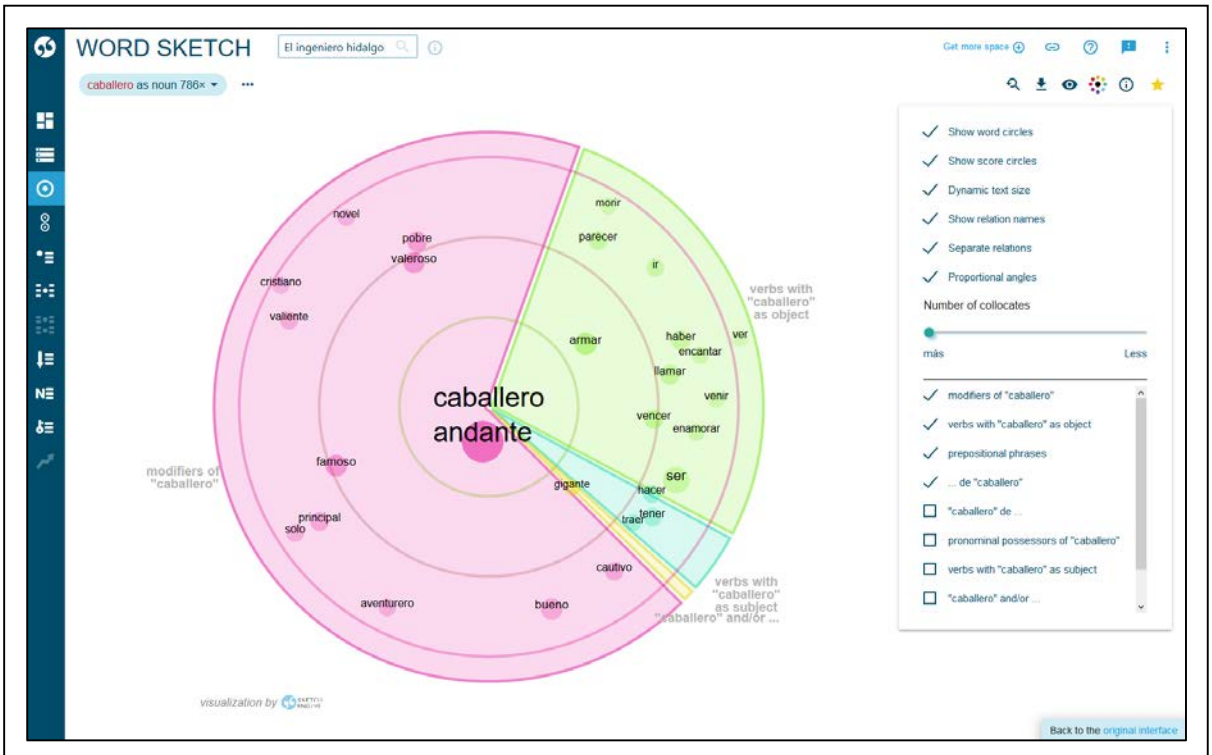

Figura 28: visualización de Word Sketch para "caballero"

Además, para cada resultado mostrado en cada una de las tablas, tenemos un icono con el símbolo **…** que permite acceder directamente a las herramientas de Concordancia y *Word Sketch* del lema buscado ("caballero" en este caso) junto con el lema que aparece en dicho resultado, además de poder acceder a la herramienta de Tesauro del lema que aparece en el resultado.

4.2. - *Diferencia Sketch*: Diferencias de uso entre palabras

Esta herramienta permite introducir **dos lemas** de búsqueda para **comparar sus patrones sintácticos y colocaciones léxicas** según ocurren en el corpus. También es posible realizar esta misma tarea usando **dos formas de un mismo lema** o un **mismo lema en dos subcorpus** de un mismo corpus.

Figura 29: menú de la herramienta Diferencia Sketch

Al clicar en esta opción en el panel de control y escoger la pestaña AVANZADO, se nos abre el menú mostrado en la Figura 29. Esta es la forma por defecto, y busca las diferencias entre dos lemas distintos, que deben compartir categoría gramatical. Si escogemos la segunda opción en la selección de arriba, podremos introducir un lema y dos formas distintas para este lema, de forma que busque las diferencias entre dichas dos formas correspondientes al mismo lema. Con la tercera opción se nos permite introducir un lema y dos subcorpus del corpus seleccionado.

Por ejemplo, para saber qué patrones son comunes a los lemas "Quijote" y "Sancho" y qué patrones son más propios de un lema o del otro, se introducen los dos en el formulario de Diferencia Sketch. Se obtiene el resultado de la Figura 30.

IMPORTANTE: los **nombres propios** se lematizan en *Sketch Engine* **en minúscula**, con lo que si se escriben con mayúscula inicial, no se encontrarán resultados.

Figura 30: resultado para Diferencia Sketch

Los colores **verde y rojo** en la Figura 30 corresponden a cada uno de los lemas introducidos. El verde es el que indica la palabra a la izquierda ("quijote") y el rojo se asocia con la palabra a la derecha ("sancho"). El grado de degradación del color se asocia con la probabilidad de compartir patrones. Cuando se degrada el color significa que la colocación es menos cercana o menos típica de la palabra

en cuestión y cuando es **blanco** es que es **común** a ambas palabras. Es decir, cuanto más intenso es el color más propia y distintiva es esta colocación para la palabra en cuestión.

Los resultados de Diferencia Sketch, además de por colores, también se muestran en los valores de las tablas. Las tablas de los patrones comunes a los dos lemas presentan 4 cifras al lado de cada una de las colocaciones. Las primeras dos cifras indican la **frecuencia de coocurrencia** con el primer lema y el segundo lema respectivamente. Las últimas dos cifras indican lo **distintivo** que es la **colocación** (***salient score***) respecto a cada lema. Las colocaciones se organizan en función del máximo de los dos índices de distinción (*salient scores*) y la coloración refleja la diferencia entre los índices (*scores*).

Esta herramienta es muy parecida a la de *Word Sketch* en cuanto a que las distintas tablas tienen encima iconos similares. Además, cada elemento de cada tabla tiene un icono **…** a la derecha que permite abrir el resultado en otras herramientas de *Sketch Engine* como son la herramienta de *Concordancia*, la de *Word Sketch* y la de *Tesauro*. Además, en la parte superior tenemos unos iconos que son los mismos que los de *Word Sketch* y tienen la misma utilidad. Entre ellas cabe destacar el icono de *Show visualization* que da una representación gráfica para las tablas de resultados, y que se muestra en la Figura 31 para la tabla de modificadores de los lemas “quijote” y “sancho”.

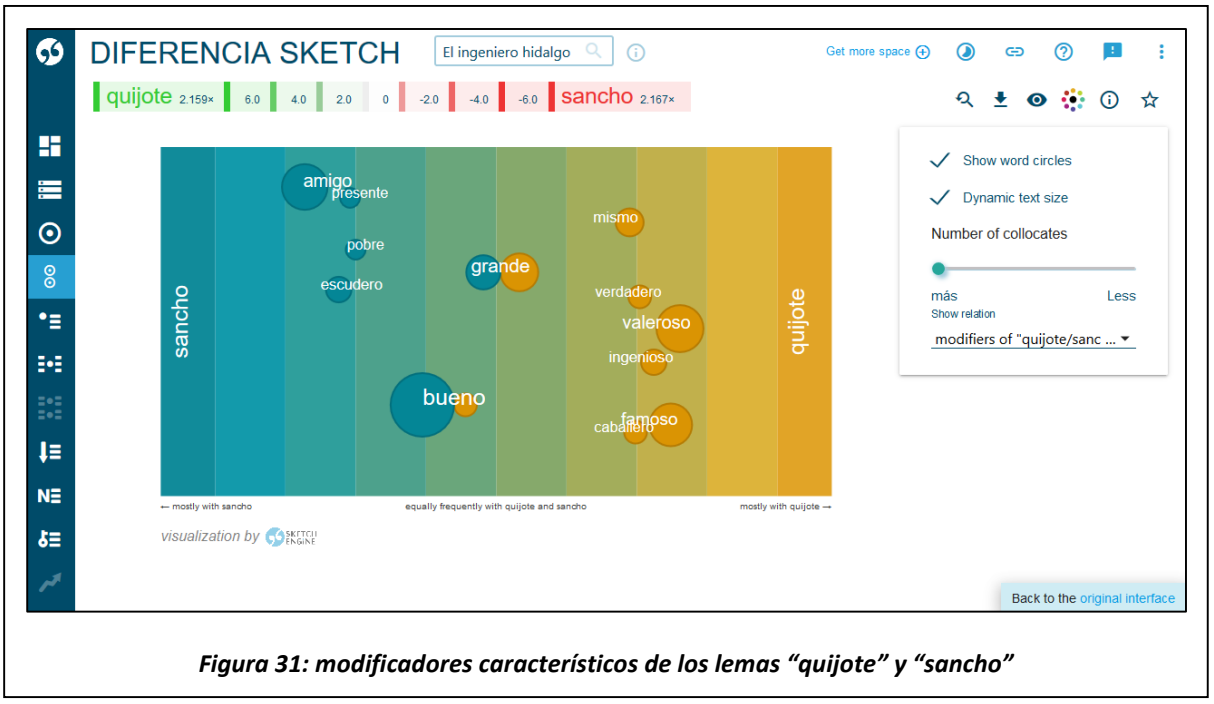

Figura 31: modificadores característicos de los lemas “quijote” y “sancho”

4.3. - *Tesauro*: Creación de un tesauro automático

Esta función calcula las palabras o **lemas que suelen aparecer con las mismas colocaciones** que una palabra dada. Basándose en estos cálculos se genera automáticamente un “**tesauro distribucional**” (*distributional thesaurus*) que recoge las palabras que aparecen en contextos similares a los de la palabra seleccionada.

Al clicar en el botón de la herramienta Tesauro, nos aparece un menú como el mostrado en la Figura 32 (como es habitual, nos centramos en la pestaña *AVANZADO*). Como puede verse, es un menú muy simple en el que únicamente se puede introducir el lema que se quiera buscar, además de escoger, si se desea, la **categoría gramatical**. Permite también decidir sobre si se agrupan los ítems similares y los

valores mínimos de puntuación necesarios para aparecer en el teasuro, así como el máximo de elementos que se mostrarán.

Figura 32: menú de la herramienta Tesauro

Por ejemplo, si se introduce la palabra "rocinante" y se genera el tesauro, como en la Figura 33, el resultado consiste en una lista de lemas (*caballo*, *asno*, *rucio*…) que tienen un comportamiento similar a la palabra "rocinante" con respecto a los patrones gramaticales y a las colocaciones.

Por lo demás, aparecen unos iconos que tienen un funcionamiento idéntico a los ya explicados para la herramienta de *Word Sketch*. Cabe resaltar el botón de *Mostrar visualización*, que crea muestra los resultados en una gráfica como la mostrada en la Figura 34.

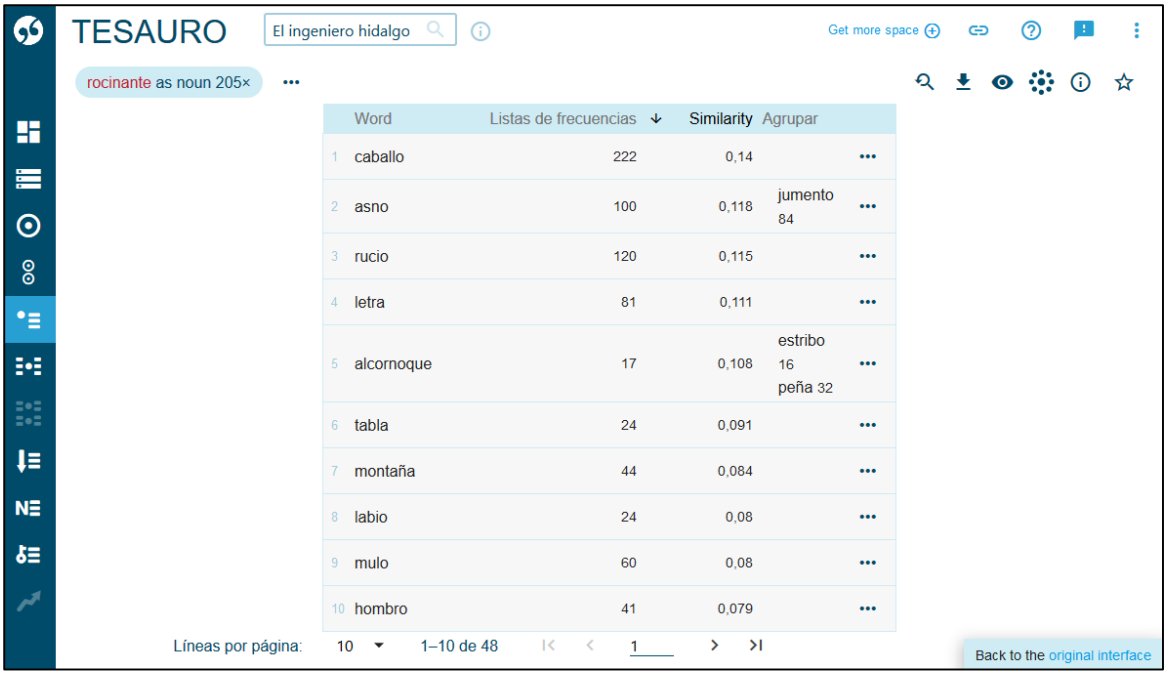

Figura 33: creación de un tesauro

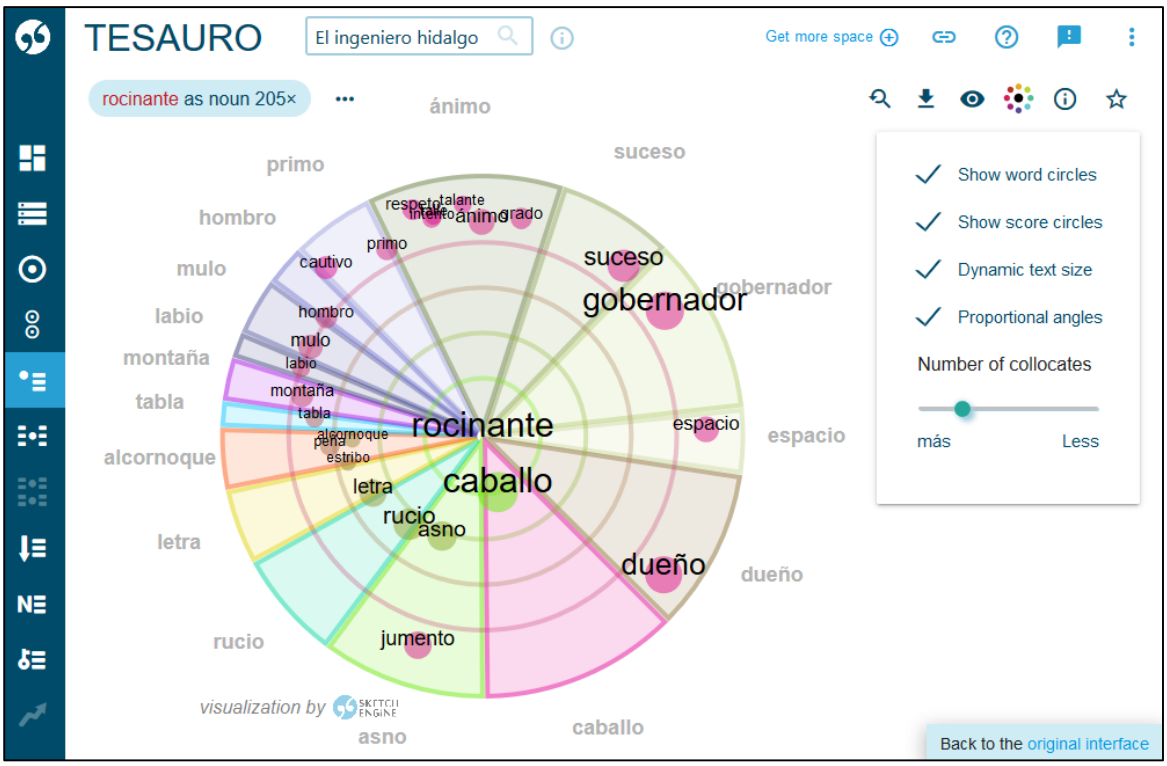

Figura 34: representación gráfica de un tesauro

4.4. - *Concordancia*: Ejemplos de uso en contexto

Al clicar en la opción *Concordancia* del menú de la izquierda, se entra en la herramienta de búsqueda de concordancias. La versión más básica de esta herramienta (pestaña *BÁSICO*) tan solo acepta un parámetro: la forma que se quiere encontrar en el texto del corpus. La versión de la pestaña *AVANZADO* es la que se muestra en la Figura 35.

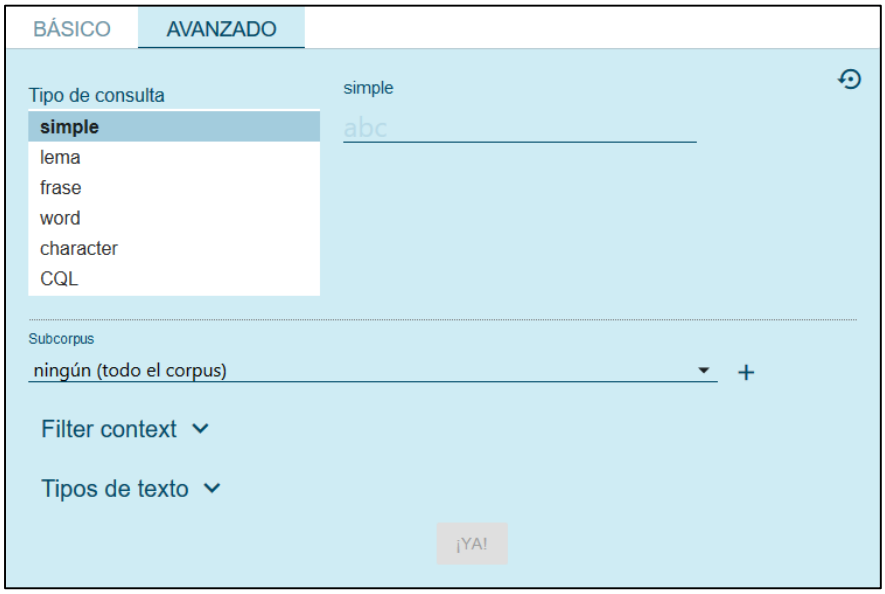

Figura 35: consulta de concordancia (AVANZADO)

Además del campo en el que escribir la propia forma cuyos contextos queremos encontrar, se nos permite limitar la búsqueda a alguno de los subcorpus. El menú da además la posibilidad de utilizar seis tipos distintos de consulta (que se explicarán en las secciones siguientes), tres tipos de filtros de contexto y la elección de los documentos del corpus que queremos utilizar para realizar la búsqueda.

El menú de ***Filter context*** permite filtrar por un contexto basado en lemas o basado en categorías gramaticales. En ambos casos, la idea es que se pueden rechazar los resultados que contengan, o no, un cierto lema o una forma de una cierta categoría gramatical en un entorno prefijado de tokens alrededor de la expresión que queramos buscar.

Debajo, clicando en ***Tipos de texto***, *Sketch Engine* permite la **elección de una parte del corpus** para hacer la búsqueda. Aparte de los subcorpus que hemos visto, hay algunos corpus que contienen **metadatos** en los archivos de texto que conforman su contenido. Uno de los metadatos más típicos es la **fecha de publicación**, pero existen otros como la **web** de la que se ha extraído el texto, o el **dominio de país** de la página web (como .es para páginas españolas, o .mx para páginas mexicanas). Sin embargo, esto no ocurre para todos los corpus, con lo que las opciones que se muestran al desplegar el menú de tipos de texto no son siempre las mismas.

4.4.1 - Tipos de consultas de concordancia

4.4.1.1 - Concordancia simple

La Figura 35 muestra la búsqueda de concordancias por defecto, que es la búsqueda simple. Tal y como se ve en dicha figura, basta con introducir una o más formas o lemas[8] en el campo ***abc***, y clicar en *¡YA!*.

Figura 36: resultado de concordancia simple para un lema

[8] En principio se pueden introducir uno o más lemas o formas flexionadas. El funcionamiento más simple es cuando se introduce una única forma flexionada: la concordancia se hará únicamente con dichas formas concretas. Pero si se introduce un lema, la búsqueda incluye a todas las posibles formas derivadas del lema. Igualmente, si se introduce más de una palabra (lema o forma), la búsqueda tratará de encontrar esa combinación de palabras, tratándolas como formas, y limitándose a ellas (aunque coincidan con los lemas).

El resultado, mostrado en la Figura 36, es una pantalla con todas las **apariciones de esta palabra en el corpus**. Por ejemplo, si se introduce el lema "molino", se obtienen todas las **líneas del corpus** (en este caso, 23) en las que aparece "molino" o "molinos". Cada línea que aparece en los resultados contiene la siguiente información:

- El **orden** de aparición **del lema/forma** (o combinación de varias formas) que se muestra. Así, el primer resultado tendrá la posición 1, el siguiente la 2, y así sucesivamente.
- Un botón que muestra un menú de **información** sobre los datos de esta aparición en concreto.
- El **orden del token** mostrado dentro de todo el corpus.
- El **contexto izquierdo** (palabras previas a la coincidencia encontrada).
- La propia **forma** que coincide con el patrón de búsqueda (destacada **en rojo**). Esta palabra es clicable, y al clicar en ella, muestra un contexto más amplio que el mostrado en el resultado.
- El **contexto derecho** (palabras posteriores a la coincidencia encontrada).

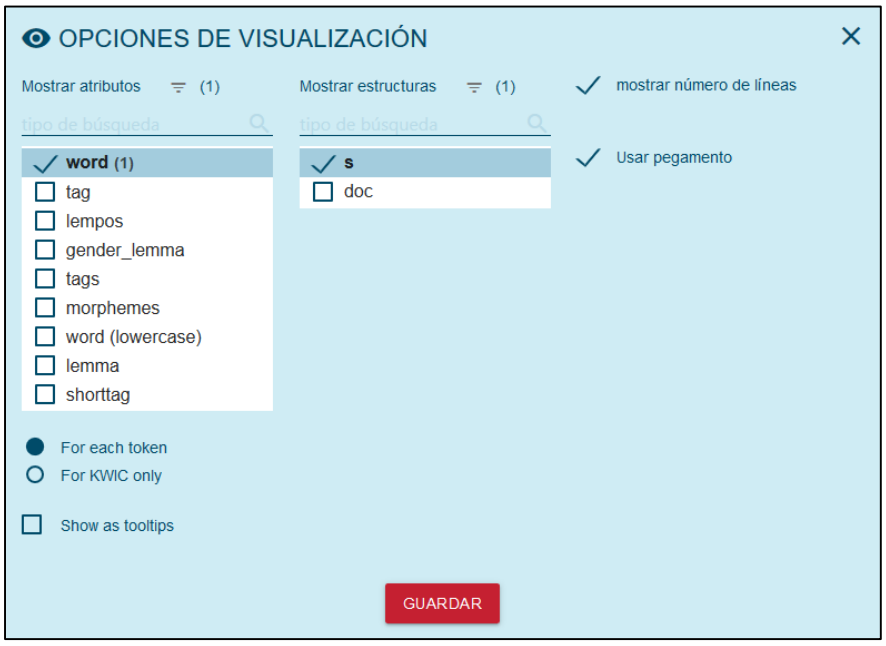

Figura 37: opciones de visualización para los resultados de concordancia

Se puede observar que, como ya hemos visto en las otras herramientas, hay una serie de iconos en la parte superior derecha que dan acceso a una serie de ajustes con respecto a los resultados mostrados. Varios de ellos son ya conocidos, pero existen unos cuantos que son propios de esta herramienta o que tienen menús distintos a los ya vistos:

- ***Deshacer última acción***: Como veremos más adelante, varias de las opciones disponibles cambiarán el resultado mostrado. Con este botón podremos deshacer el último cambio realizado.
- ***Opciones de visualización***: Ya se ha visto previamente este botón, y esta vez muestra un menú como el que aparece en la Figura 37. En la Figura 36, cada una de las palabras del corpus aparecía sin ninguna información extra. Pero se puede mostrar cada palabra acompañada de ciertos datos como:
    - La **etiqueta sintáctica individual** con la que se etiqueta cada una de las formas.

- El **lema-categoría** del que deriva. Por ejemplo, la forma "amaneceres" deriva del lema-categoría *amanecer-Nombre*, mientras que "amanecerá" deriva de *amanecer-Verbo*.
- El **lema manteniendo el género** de la forma. Para palabras con variación en género y que aparecen en género femenino, se muestra el lema en femenino sin pasarlo al masculino como suele ser habitual. Así, "hermosas" se lematizará como *hermosa*, mientras que "hermosos" se lematizará como *hermoso*.
- Las **etiquetas sintácticas múltiples** que pueda tener la forma. Ocurre en algunas formas que pueden etiquetarse con más de una etiqueta (en español esto ocurre en las formas verbales con clíticos y las contracciones). Así, la forma "decírselo", tiene una etiqueta individual de VMN0000, pero una etiqueta multiple en la que cada clítico tiene su propia etiqueta: VMN0000, PP3CN00, PP3MSA0.
- Los **morfemas**. En realidad, esto se aplica al mismo tipo de formas que en el punto anterior, de manera que para la forma "del" muestra su división en lo que *Sketch Engine* denomina morfemas: de, el.
- La propia **palabra en minúscula**.
- El **lema** del que deriva la forma.
- La **etiqueta corta**. Esta etiqueta consiste simplemente en el primer carácter de la etiqueta completa (que puede tener hasta un máximo de 7 caracteres para verbos y nombres en español), que es el que identifica la categoría gramatical.

También se puede decidir si esta información se muestra únicamente para la palabra para la que hemos solicitado la concordancia, o para todas las palabras incluyendo ambos contextos. Además, dicha información puede mostrarse directamente en pantalla bajo cada una de las formas, o permanecer oculta hasta que ponemos el puntero del ratón sobre la forma.

Por último, se puede ocultar el número de orden que aparece a la izquierda de cada línea.

- ***Tomar una muestra aleatoria***: Muestra una cantidad limitada de resultados tomándolos aleatoriamente de entre todos los resultados existentes.
- ***Ordenar***: El orden por defecto en el que se muestran los resultados es el de aparición en el corpus. Con esta opción se permite ordenar (alfabéticamente) no solo según la palabra central sino también según palabras del contexto situadas hasta tres posiciones antes o después.
- ***Filtrar***: Con esta opción se permite crear reglas de filtrado basadas en la aparición de una forma, lema, carácter… situado en el contexto de hasta cinco palabras previas o posteriores.
- ***Buenos ejemplos para diccionario***: Se muestran primero aquellos ejemplos de uso que se consideran buenos ejemplos para diccionario[9]. Se considera que una aparición de una forma es un buen ejemplo de uso de dicha forma, si aparece en una frase de 10-25 palabras en la que no aparecen otras palabras raramente usadas, a ser posible sin anáforas, si la forma aparece en una colocación típica…
- ***Frecuencia***: Se muestran los datos de frecuencia de aparición (apariciones por millón de formas) de la palabra clave buscada, así como de las palabras que la rodean en el corpus. Se puede calcular la frecuencia según la etiqueta morfosintáctica, la forma o los lemas. En el

[9] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. y Suchomel, V. (Eds.) 2013. The tenten corpus family.
Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus.

ejemplo mostrado en la Figura 38 se muestran las diez etiquetas con mayor frecuencia para el lema “comer”.

Figura 38: frecuencias de aparición de las distintas etiquetas morfosintácticas para el lema “comer”

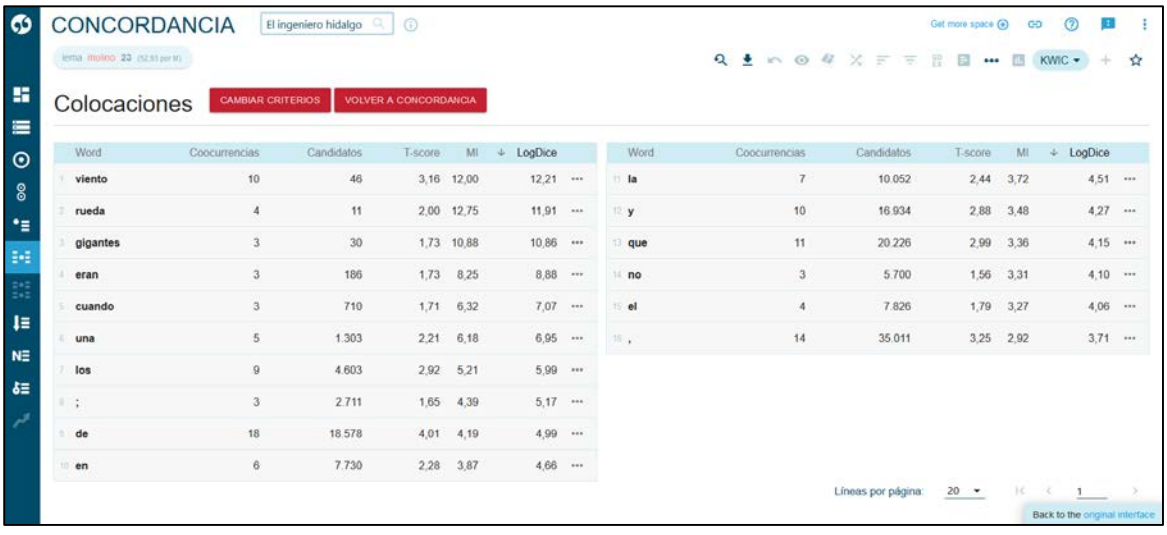

Figura 39: colocaciones para la palabra “molino”

- ***Colocaciones***: Se ofrece un menú en el que se puede escoger el entorno alrededor de la palabra clave en el que se buscan otros lemas que puedan formar potencialmente colocaciones con la palabra dada. Dado que se busca palabras que coaparecen

frecuentemente con la palabra clave, se intenta eliminar las palabras muy comunes (como preposiciones o determinantes) penalizándolas. Para ello se dan varias fórmulas distintas para dicha penalización, utilizando distintos modelos estadísticos. El resultado para "molinos" se muestra en la Figura 39, utilizando los modelos por defecto (T-score, MI y logDice) y un entorno de ±5 palabras. Puede apreciarse como la colocación más habitual es "viento" usando el modelo logDice, mientras que es "rueda" si se usa MI y un sorprendente "de" cuando se usa el modelo T-score.

- ***Distribución de ocurrencias en el corpus***: Con este botón del menú, se puede obtener una "radiografía" de la aparición del término buscado en el corpus. Por ejemplo, podemos ver en qué partes del corpus aparece más a menudo el lema "comer". El resultado, mostrado en la Figura 40, indica que el lema está muy repartido a lo largo del corpus, si bien existen unas ciertas partes donde este lema es especialmente abundante. Puede cambiarse la granularidad: en el ejemplo dado, cada barra incluye un 0,5% del corpus, pero dicho valor puede oscilar entre el 0,1% y el 10%. Además, cada barra es clicable, y al hacer click se nos muestran los resultados de concordancias dentro de esa parte del corpus.

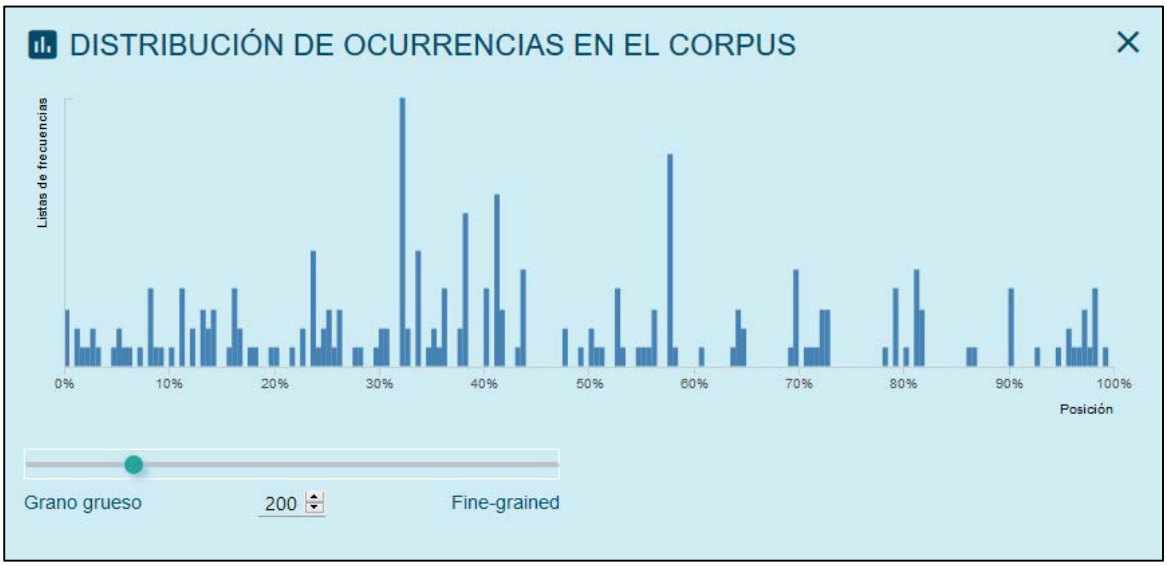

Figura 40: colocaciones para la palabra "molino" en el corpus de El Quijote

- ***KWIC***: Es el acrónimo de *KeyWord In Context*. Con este menú podemos decidir si el entorno que se nos muestra en los resultados es el correspondiente a un cierto número de palabras previas y posteriores (cuya longitud se fija en las opciones de visualización explicadas más arriba), o si el entorno es simplemente la frase completa en la que aparece.
- ***Crea subcorpus***: Con esta opción, se puede crear un subcorpus formado por las frases o documentos que contengan resultados de concordancia.

4.4.1.2 - Concordancia de lema

Para utilizar este tipo de consulta de concordancia se debe seleccionar ***lema*** en el menú de la Figura 35. Al hacerlo, el menú nos permite introducir un lema y también una **categoría gramatical** (cuando tiene más de uno). Así, la palabra "sobre" es un lema, que puede hacer referencia a una preposición o

a un nombre[10], o "anochecer" puede ser un verbo o un nombre. Si no se especifica la categoría, se escoge la más frecuente (en caso de que en la consulta simple se introduzca un lema, dicha consulta sería equivalente a esta consulta de concordancia de lema).

Por lo demás, este tipo de búsqueda es muy similar a la explicada en la subsección anterior y el tipo de resultados y opciones de visualización son los mismos.

4.4.1.3 - Concordancia de frase (sintagma)

La opción ***frase*** de la Figura 35 nos permite hacer una búsqueda por **sintagmas**. Se pueden introducir varias palabras y se obtendrán las apariciones de exactamente dichas formas (se consideran formas flexionadas aunque coincidan con lemas), exactamente en el orden dado, y sin ninguna otra palabra entre medias. Al no dar ninguna información de categoría gramatical, *Sketch Engine* devuelve todas las apariciones de dicha combinación de palabras independientemente de su categoría gramatical. Si se busca "la vista", se devolverían resultados para la combinación de artículo + nombre ("ha perdido la vista"), pero también para la de pronombre + verbo ("necesito que la vista").

Este tipo de búsqueda puede conseguirse también si en la búsqueda simple se introducen los (varios) términos que se pretende encontrar. Por lo demás, la forma de mostrar los resultados es la misma que en los modos de búsqueda descritos anteriormente.

4.4.1.4 - Concordancia de forma

La opción ***word*** es similar a las anteriores, con la diferencia de que limita el número de resultados a la **forma concreta** dada (aunque coincida con un lema) pero a la vez, da la capacidad de elegir el **tipo de palabra** (categoría gramatical) que se desea. Así, se puede buscar el término "bajo" como adjetivo, en cuyo caso devolvería concordancias únicamente con dicha forma (solamente en masculino singular), o como preposición.

4.4.1.5 - Concordancia de caracteres

Para este tipo de búsqueda (usando la opción ***character*** en el menú de *Query type*) se devuelven los resultados al buscar la ocurrencia de la **cadena de caracteres** dada, aunque **no se trate de una palabra completa**. Para los tipos de búsquedas vistos hasta ahora, si se busca por "vista" solo se devolverán resultados para esta palabra en concreto. Si esta misma búsqueda se utiliza en el campo *character*, también se devolverán resultados para otras palabras como "revista" o "vistazo". Igualmente, si se busca por ejemplo el término "ábamos", se nos devolverán las terminaciones de las formas verbales que sean de primera persona del plural del pretérito imperfecto de indicativo de verbos de la primera conjugación. No obstante, existen formas mejores de conseguir este resultado.

4.4.1.6 - Concordancia CQL

Este es el tipo de búsqueda más completo, ya que permite definir de forma muy precisa los resultados que se buscan. Si se escoge la opción ***CQL*** (de *Corpus Query Language*), el texto a introducir en el campo *CQL* será una expresión compuesta por elementos del tipo **[atributo="valor"]**, donde el valor puede contener además expresiones regulares. Existen varios atributos, entre los que destacan ***lemma*** (lema), ***word*** (forma flexionada) y ***tag*** (etiqueta morfológica), pero en realidad existen tantos atributos

[10] También podría considerarse como una forma flexionada del verbo "sobrar", pero en dicho caso no sería un lema, sino una forma.

como los que se han visto en las opciones de visualización de los resultados, tal y como puede verse en la Figura 41.

Si bien esta búsqueda es la que proporciona mayor libertad a la hora de escoger qué concordancias en concreto queremos encontrar, a cambio tiene cierta complejidad en su sintaxis ya que hace uso de expresiones regulares básicas y de etiquetas morfosintácticas. Debido a ello, *Sketch Engine* proporciona un motor de creación para estas expresiones, además de incluir un pequeño tutorial de vídeo acerca de su uso tal en el propio menú, tal y como puede observarse en la Figura 41. Por ejemplo, la expresión:

```
[word=".*[aeií]r" & tag="[^V].*" ]
```

devuelve todas las palabras que acabaran en –ar, -er, -ir, o –ír que no sean verbos (como “lugar”, “mujer” o “Sir”). Igualmente, la expresión

```
[morphemes="meter,.*"] [shorttag="S" & word!="en"] [tag="D[ADP].*"] []? [lemma="asunto"]
```

devuelve como resultados expresiones que:

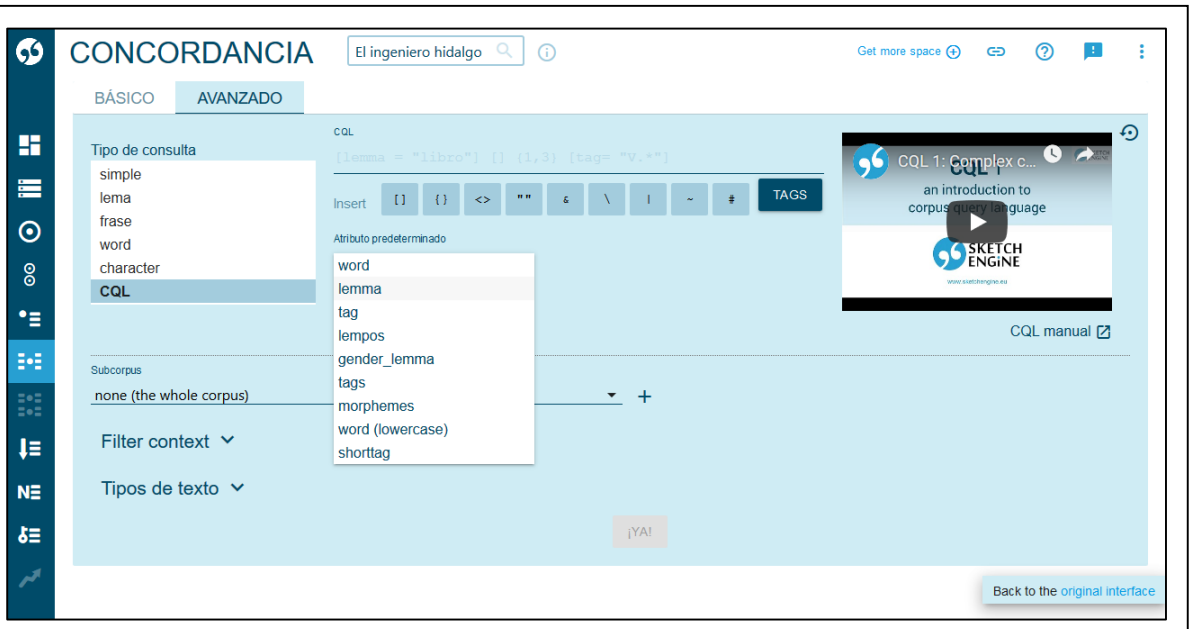

Figura 41: consulta de concordancia usando expresiones CQL

- Comiencen con el lema “meter” (que solo puede ser verbo) conjugado en cualquiera de sus formas y estén seguidos de al menos un clítico.
- Tengan a continuación una preposición que no sea “en”.
- Continúe con un determinante que sea un artículo, posesivo o demostrativo.
- Pueda llevar un token de cualquier tipo a continuación.
- Termine con una forma flexionada del lema “asunto”.

Con esta expresión capturaríamos expresiones como “métase con el malditos asuntos” o “metiéndolo a ese asunto”, que indicarían variaciones de la colocación típica “meterse en mis asuntos”. Estas búsquedas pueden tardar varios minutos en realizarse si el corpus es muy grande.

4.5. - *Listas de palabras*: frecuencias de aparición

Para obtener la lista de palabras de un corpus junto con las frecuencias de cada una, se puede clicar en ***Lista de palabras***, que aparece en la parte superior del menú izquierdo. En el menú que aparece en la Figura 42, puede apreciarse que hay diferentes opciones para obtener la lista de palabras y su **frecuencia**. La más básica se corresponde con *word*, que hace **recuentos de apariciones** de palabras diferenciando por sus formas. Sin embargo, existen multitud de variantes, comenzando por el recuento a nivel de lemas, o de etiquetas sintácticas, por ejemplo, además de las distintas categorías gramaticales.

Puede hacerse el recuento de los *tokens* que empiecen o terminen por una cierta cadena de caracteres, o que la contengan. También pueden usarse **expresiones regulares** para definir las ocurrencias que entrarán en el recuento, y puede incluirse también una lista de palabras concretas que no deben incluirse en el recuento aunque cumplan el resto de condiciones. Incluso pueden eliminarse del recuento los *tokens* que no sean palabras (como símbolos y signos de puntuación).

Por último, existe la posibilidad de seleccionar cómo se agrupan los recuentos. Por defecto se muestra una lista simple, en la que se tiene una única columna de lema, forma, etiqueta… y otra columna con las frecuencias. Sin embargo, si se usa ***Mostrar como*** es posible mostrar más de una columna, con lo que se puede mostrar, por ejemplo, un listado de formas, con su correspondiente etiqueta morfosintáctica y el lema, seguido de su frecuencia.

Figura 42: opciones de recuento de aparición de palabras

4.6. - *N-gramas*: extracción de expresiones multipalabra

Con esta herramienta se puede analizar el corpus para encontrar las expresiones multipalabra más típicas. Al clicar el botón de N-gramas[11], se muestra el menú que aparece en la Figura 43. En el menú se puede escoger el ***Tamaño de N-gramas***, especificando un rango de longitud de las expresiones multipalabra que se buscarán (por defecto 2 o 3). Es posible buscar recuentos de formas, lemas, etiquetas… hasta los ocho tipos que ya se han visto en las opciones de visualización de la Figura 37. También se puede restringir la búsqueda de la expresión multipalabra indicando las letras por las que debe empezar, usando expresiones regulares, o indicando las palabras con las que debe empezar o acabar.

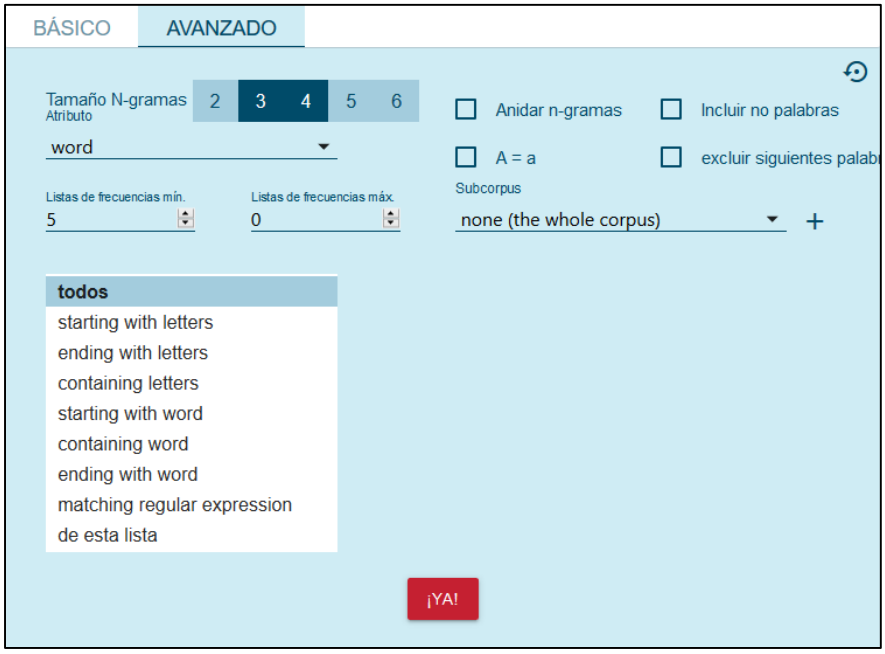

Figura 43: menú de la herramienta N-gramas

También se da la opción de anidar los resultados obtenidos (de forma que se muestren juntas las expresiones que empiecen por las mismas palabras, independientemente de que por su frecuencia deberían aparecer más tarde), incluir *tokens* que no sean palabras, diferenciar o no entre mayúsculas o minúsculas, y excluir términos multipalabra que incluyan alguna forma de las incluidas en una lista.

Tras clicar en el botón *¡YA!* se nos muestra la lista de resultados. En la Figura 44 se muestra el listado de las 50 expresiones de cuatro lemas más frecuentes en el corpus de El Quijote.

[11] Un n-grama es simplemente un conjunto de n palabras que aparecen de forma consecutiva en el corpus. Cuando n es igual a 2 se habla de bigramas, y cuando es igual a 3, de trigramas. Para números de n superiores a 3 (cuyo uso no es muy común), se habla de 4-gramas, 5-gramas… y así sucesivamente.

Figura 44: menú de la herramienta N-gramas

4.7. - *Palabras clave*: extracción de palabras clave y términos multipalabra

Por último, se tiene la herramienta ***Palabras clave*** que hace un análisis a nivel del corpus completo (tiene que ser un corpus creado por el usuario) y extrae de él las **palabras clave** más características, así como los **términos multipalabra** más propios de dicho corpus. Esto lo hace analizando el texto del corpus del que se quieren extraer las palabras clave, y **comparando** dicho análisis con un **corpus equilibrado**, que representa múltiples temas y múltiples variedades de la lengua. Así, se identifican aquellas palabras o grupos de palabras que resultan frecuentes en el corpus que está siendo analizado, pero que no lo son tanto en un corpus mayor y con contenidos más variados.

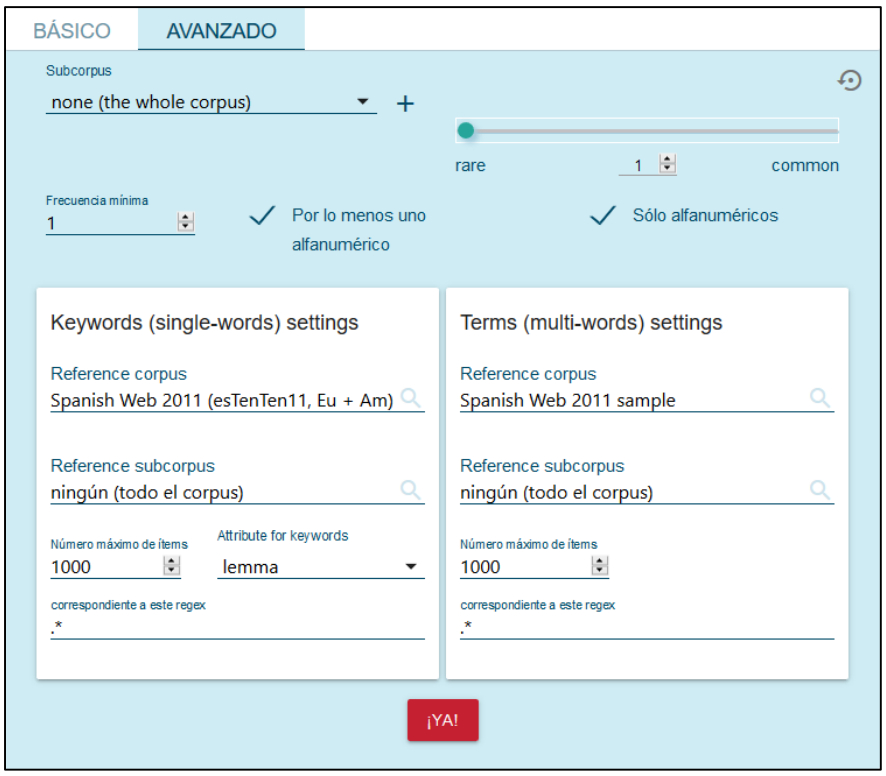

Figura 45: menú de la herramienta Palabras clave

Al clicar en el botón de esta herramienta se nos muestra el menú que aparece en la Figura 45. Entre otras cosas, en él se puede escoger cuál es el corpus de referencia frente al que se compara nuestro corpus privado, y algunas características de los términos que se extraerán.

Tras pulsar en el botón *¡YA!* se obtienen los resultados, que se muestran en la Figura 46. Como puede apreciarse, al estar analizando un corpus que consiste en las dos partes de El Quijote, las palabras clave son, fundamentalmente, nombres propios usados en el universo de El Quijote, junto con conceptos típicos como “caballero andante” o “libro de caballería”.

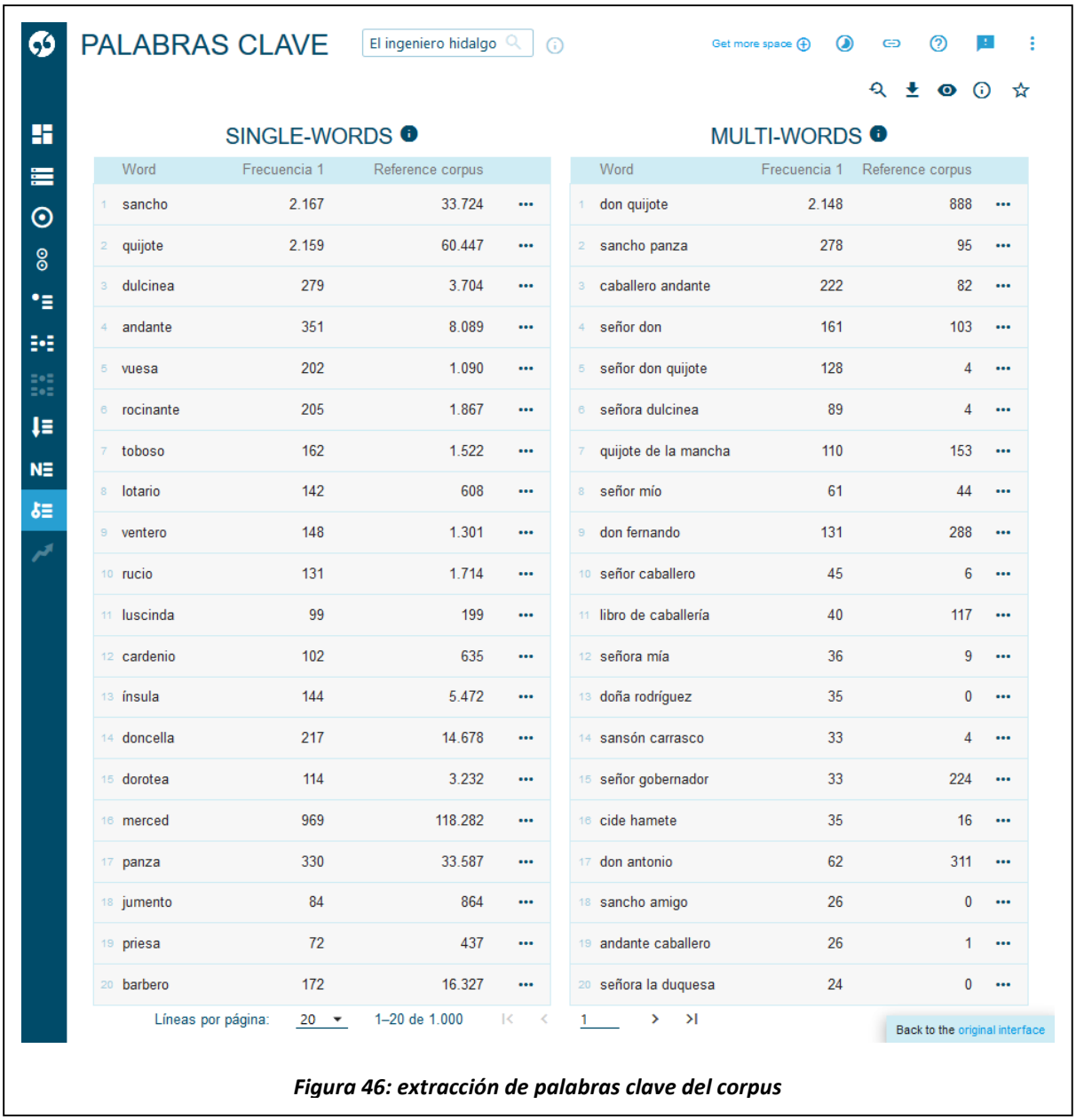

Figura 46: extracción de palabras clave del corpus

4.8. - *OneClick Dictionary*: creación de un diccionario automático

Con esta herramienta, *Sketch Engine* es capaz de crear un esbozo de diccionario utilizando para ello las concordancias en el corpus. Esta herramienta, que necesita una cuenta en la web https://www.lexonomy.eu/, hace un análisis gramatical del corpus para extraer atributos de términos e inferir su significado.

Como su propio nombre indica, basta con clicar en la herramienta para que se genere el diccionario utilizando el corpus seleccionado.

4.9. - *Corcondancia paralela*: expresiones equivalentes en dos lenguas

Con esta herramienta se utiliza un corpus paralelo en varias lenguas para extraer las expresiones equivalentes en dos lenguas. El menú es similar al de la herramienta de *Concordancia*, con la diferencia de que se realiza en dos lenguas. Se debe escoger un término en una lengua origen, y una lengua

destino. Una vez escogida la expresión que se quiere traducir a otra lengua, se obtiene un resultado como el mostrado en la para la palabra española “toro”y su equivalente en búlgaro.

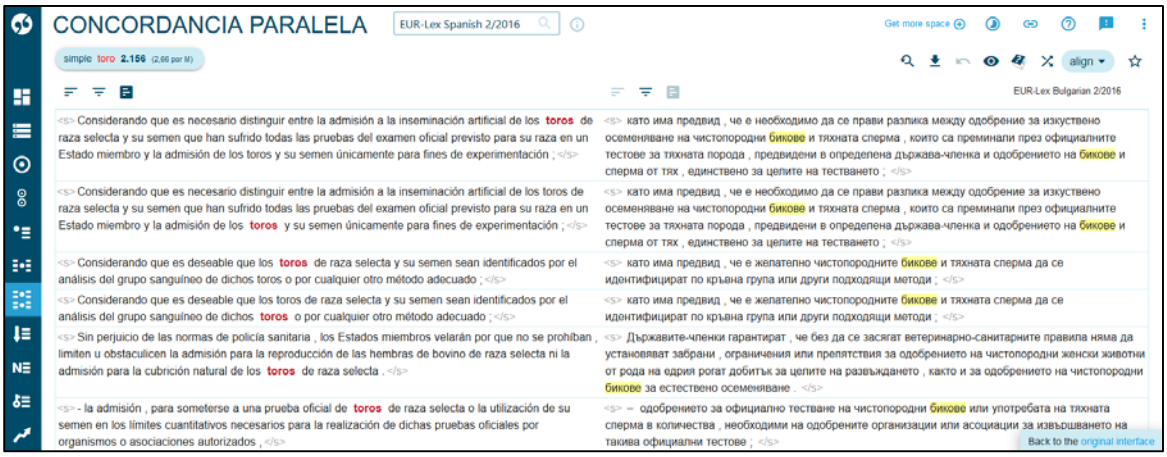

Figura 47: corcondancia paralela entre la palabra española “toro” y su correspondiente en búlgaro

4.10. - *Tendencias*: variación en el uso de palabras a lo largo del tiempo

Para poder usar esta herramienta es necesario tener seleccionado un corpus cuyos documentos estén etiquetados según su fecha de creación. De esta manera, *Sketch Engine* es capaz de comprobar qué palabras se usan más o se usan menos en los documentos más recientes en relación con los más antiguos. El menú de la herramienta permite decidir si mostrar la tendencia para formas, lemas… o cualquiera de las ocho opciones que se muestran en la Figura 35. El resultado para el corpus EUR-Lex en español aparece en las figuras a continuación: la Figura 48 muestra las palabras que se usan menos en documentos más recientes, mientras que Figura 49 se muestran las que se usan más.

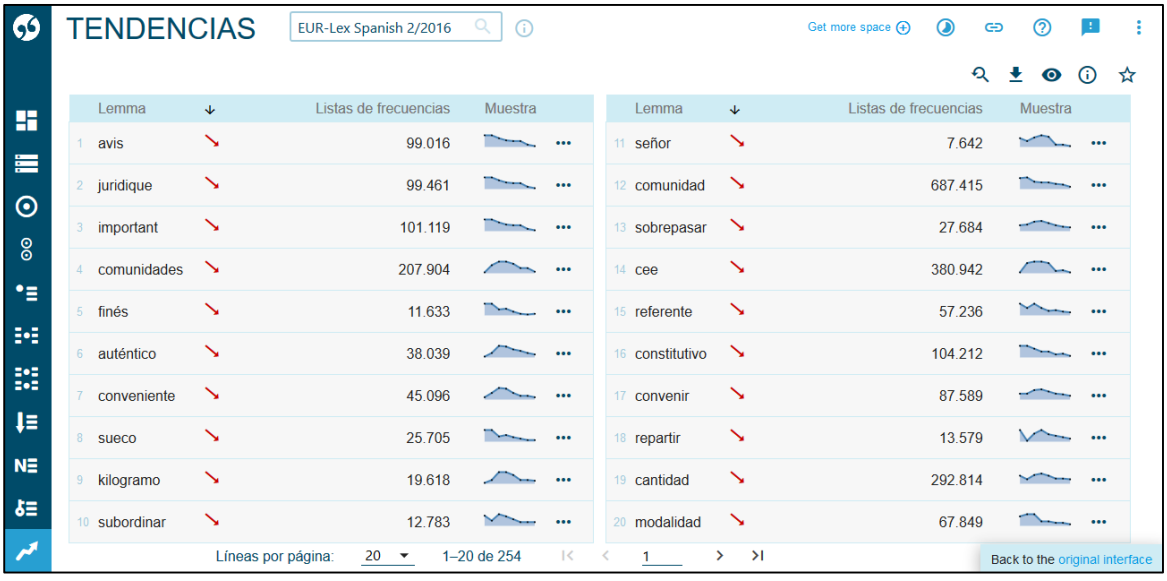

Figura 48: principales tendencias negativas (palabras en desuso)

Figura 49: principales tendencias negativas (palabras de moda)