Elena Battaner Moro & Cristina V. Herranz-Llácer
& Ana Segovia Gordillo

Corpus y herramientas digitales para el estudio de la historiografía lingüística hispánica*

1. Metodología: herramientas digitales (corpus y software de trabajo)

1.1 El corpus de estudio: BiTe_Corpus

El material sobre el que se ha confeccionado el corpus denominado BiTe_Corpus[1] es el conjunto de los resúmenes que formaban parte de las fichas bibliográficas de la *Bibliografía Temática de la Historiografía Lingüística Española. Fuentes secundarias* (BiTe) (Esparza *et al*. 2008). En un principio, este material constaba de un total de 296 029 palabras y 31 469 vocablos (apariciones únicas de palabras) distribuidos en 2298 resúmenes —ordenados según los capítulos de BiTe—. Cuando se procedió a su edición, se desarrolló un trabajo detallado, lento y minucioso que llevó más de un año. Se tomaron como punto de partida los criterios de edición publicados por Samper Padilla (1998), usados tradicionalmente para el estudio de la disponibilidad léxica; además, siempre con la idea de observar los detalles y singularidades del corpus, a la hora de tomar decisiones sobre la edición, se trató de mantener una postura conservadora (frente a la uniformadora) (Fernández Juncal 2013).

El corpus comenzó a editarse en un documento de texto en formato Word. Como se puede apreciar en la Figura 1, este documento inicial, extraído de la base de datos *FileMakerPro*, numeraba cada una de las líneas del resumen; así pues, una de las primeras tareas consistió en eliminar dichos números y eliminar los resúmenes que no estuvieran publicados en lengua española —en la BiTe hay trabajos en inglés, alemán, francés, italiano o catalán, entre otros—. En la Figura 1 se muestra una imagen del texto en bruto.

* Grupo de investigación de alto rendimiento LiyNMedia, Universidad Rey Juan Carlos.

[1] La obtención de estos resúmenes se hizo gracias al Dr. D. Miguel Ángel Esparza Torres, a quien desde estas líneas queremos agradecer su colaboración, pues extrajo de la base de datos los resúmenes de los trabajos controlados (cerca de 3200 registros).

Artículo recibido el 21/07/2021 y aceptado el 01/09/2021

245 "La época que va desde Elio Antonio de Nebrija hasta Sebastián de Covarrubias merece con todo derecho el título de 'Siglo de la lexicografía'. Aunque sus comienzos son bastante anteriores a la creación de la Herzog August Bibliothek, sus fondos nos permiten estudiarlo y describirlo de forma satisfactoria. La gramática, a pesar de su brillante surgimiento a cargo de Nebrija, queda relegada durante esta época a un triste segundo plano. Hacia la segunda mitad del siglo vuelve tímidamente a aparecer, si bien orientada hacia finalidades de tipo práctico que apenas merecen ser calificadas de científicas. La situación no cambia hasta los años que preceden al siglo XVII, es decir con la Gramática de Oudin, hecho que se desprende con claridad de los fondos de la Herzog August Bibliothek. Los tratados de índole más general acerca del origen y la historia del castellano, aparecidos en España a partir de los mediados años ochenta del siglo XVI, son los únicos de la época a que nos referimos (antes de 1611) que no pueden localizarse en Wolfenbüttel."

246 "La lexicographie de l'espagnol commence au Moyen Âge, avec un nombre assez réduit de glossaires. Elle ne prend son véritable essor que vers la fin du 15e siècle, avec les dictionnaires latin-espagnol et espagnol-latin de Nebrija. Suivent les dictionnaires latin-espagnol de Maître Rodriguez et de Jiménez Arias. A partir des années 1570, c'est le Calepin, enrichi d'une glose espagnole, qui commence à viser le même public que ces petits dictionnaires polyglottes destinés aux commerçants de l'époque et qui sont publiés pendant tout le siècle. Vers 1570 également, dans une outre série de dictionnaires, le latin est remplacé par une langue non-classique; le dictionnaire bilingue moderne est né. Quarante ans plus tard, Covarrubias publiera son «Tesoro», premier véritable dictionnaire monolingue de l'espagnol. Quant aux lieux d'impression de tous ces dictionnaires, les Pyrénées semblent avoir constitué une véritable barrière, et cela dans les deux sens. Seul Nebrija a pu la franchir."

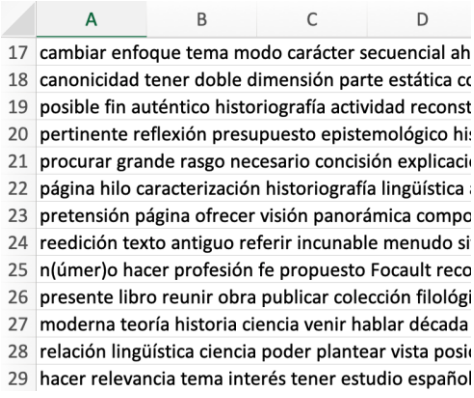
247 "Mi propósito es darles aquí una visión panorámica de lo que se ha hecho en este campo en aquellos tiempos. No hago, pues, una diferencia entre inventarios lexicográficos que toman como punto de partida el español y los que se sirven del español para explicar otros idiomas, sino que hablaré de toda obra lexicográfica española siempre que en ella el español aparezca de cualquier manera que sea, ordenada en voces de entrada o como glosa explicativa."

Figura 1. Visión del corpus inicial en Word. Fuente: elaboración propia.

El siguiente paso fue la eliminación de todos los signos de puntuación y, seguidamente, las denominadas *stopwords* clásicas: preposiciones, conjunciones, determinantes, pronombres y algunos verbos auxiliares. Somos conscientes de que esto ha podido marcar el devenir "semántico" del corpus, pero es una decisión que, en un momento dado, hay que tomar. En este caso, a pesar de ser palabras frecuentes porque aparecen de forma constante en los textos, no tienen significado léxico, sino gramatical y son definidas como palabras vacías (Cuartero Sánchez 2012). Por tanto, se decidió confeccionar un corpus en el que básicamente quedaran sustantivos, adjetivos, verbos y algunos adverbios. Como señalamos más arriba, se decidió no modificar el orden de aparición de las palabras dentro de cada uno de

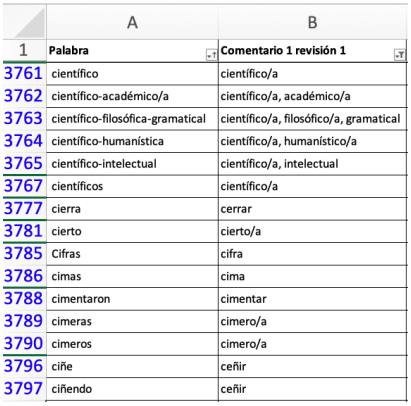
los resúmenes y se mantuvo el orden de cada resumen dentro de los quince capítulos de BiTe[2]. Llegados a este punto, se determinó que, para ser más sistemáticos y agilizar el proceso de edición, trabajaríamos con el programa *Microsoft Excel*.

En primer lugar, se copió el corpus con los cambios señalados arriba en un documento Excel, de forma que cada uno de los resúmenes estuviera delimitado en una celda de este programa (v. Figura 2).



	A	B	C	D
17	cambiar enfoque tema modo carácter secuencial ah			
18	canonicidad tener doble dimensión parte estática cc			
19	posible fin auténtico historiografía actividad reconst			
20	pertinente reflexión presupuesto epistemológico his			
21	procurar grande rasgo necesario concisión explicaci			
22	página hilo caracterización historiografía lingüística			
23	pretensión página ofrecer visión panorámica compo			
24	reedición texto antiguo referir incunable menudo si			
25	n(úmer)o hacer profesión fe propuesto Focault reco			
26	presente libro reunir obra publicar colección filológi			
27	moderna teoría historia ciencia venir hablar década			
28	relación lingüística ciencia poder plantear vista posi			
29	hacer relevancia tema interés tener estudio español			

Figura 2. Corpus parcial en Excel. Fuente: elaboración propia.



	A	B
1	Palabra	Comentario 1 revisión 1
3761	científico	científico/a
3762	científico-académico/a	científico/a, académico/a
3763	científico-filosófica-gramatical	científico/a, filosófico/a, gramatical
3764	científico-humanística	científico/a, humanístico/a
3765	científico-intelectual	científico/a, intelectual
3767	científicos	científico/a
3777	cierra	cerrar
3781	cierto	cierto/a
3785	Cifras	cifra
3786	cimas	cima
3788	cimentaron	cimentar
3789	cimeras	cimero/a
3790	cimeros	cimero/a
3796	ciñe	ceñir
3797	ciñendo	ceñir

Figura 3. Lista parcial de palabras únicas en Excel. Fuente: elaboración propia.

Y, en segundo lugar, gracias al lenguaje de programación VBA (*Visual Basic* para Aplicaciones), se obtuvo el listado de palabras únicas (Figura 3) que se fue editando según los criterios que explicamos más adelante —a excepción de los casos en los que la palabra no requería cambios, en cuyo caso se marcaba la revisión con un guion—. En este punto también se eliminaron todas aquellas palabras que estuvieran escritas en un idioma distinto al español, aunque pudieran estar contenidas en un resumen escrito en español. Así, todas aquellas palabras provenientes del inglés, francés, alemán, italiano o latín fueron excluidas.

Algunas de las decisiones que se tomaron con respecto a la edición del corpus son las siguientes:

- Cada unidad léxica es una única palabra, a excepción de antropónimos y topónimos. Para estos casos, se empleó, por un lado, el guion bajo (_) para la unión de cada uno de los componentes del nombre; y, por otro, los paréntesis, para mostrar la presencia o ausencia de elementos de la

[2] Es fundamental señalar esta circunstancia porque explicará determinados comportamientos a la hora de analizar algunos resultados. Esta característica afecta específicamente al análisis del corpus con *Voyant*, como veremos en la sección 2 de este artículo.

palabra. Por ejemplo, autores como Vicente Salvá quedan registrados como *(V(icente)) Salvá*. Esta forma final representa, por tanto, todas aquellas maneras de citar al autor en los resúmenes de la BiTe: *Vicente Salvá, V. Salvá* o *Salvá*.

- El uso del paréntesis se empleó también para la unificación de vocablos que muestran distintas formas: por ejemplo, las palabras *hispánico*, *hispánica* o *hispánicas* se unificaron en *hispánico/a* o Gerónimo Mendieta y Jerónimo Mendieta quedó como *(G/Jerónimo) Mendieta*. Esta decisión se ha mantenido siempre que el añadido no supusiera una ampliación o limitación del significado como ocurre con *Estados_Unidos* y *Estados_Unidos_Mexicanos*, ya que se nombran países diferentes.
- No se mantuvieron las mayúsculas derivadas de la puntuación, de forma que solo se conservaron las mayúsculas de antropónimos, como en *(C(ésar)) Oudin*; topónimos (*Chile*, *España, Asia*, entre otros) y siglas no desarrolladas en el corpus (por ejemplo, *ALEICan*).
- Los verbos se registraron en infinitivo; así, en el corpus, la palabra *suprimir* contendría las voces *suprime, suprimido*, *suprimiendo* y *suprimió*, que son las formas de este verbo que aparecen en los resúmenes.
- Los sustantivos que presentaban variabilidad genérica y los adjetivos de dos terminaciones se transcribieron siguiendo las siguientes normas: a) si solo aparecían en femenino, se mantuvo la forma femenina (*adulta* o *divulgadora*); b) si solo aparecían en masculino, se conservó la forma masculina (*cartesiano* o *gaditano*); y c) si aparecían en ambas formas, se incluyó el desdoblamiento con barra unificándolos en una única entrada, tal y como ocurre con *valenciano/a*, *santo/a*, *riguroso/a*, etcétera.
- Los sustantivos que aparecen en plural se cambiaron a singular, siempre que esa palabra estuviera previamente en singular: *ámbito, ambivalente*, *andalucismo*, etc.
- Se mantuvo el polimorfismo gráfico siempre que fue posible empleando el corchete para indicar la aparición de ambas voces indistintamente. Por ejemplo, se registró *qu[e/i]chua* para *quechua*, *quichua*, *quechuas* y *quichuas* o *eus[k/qu]era* para *euskera* y *eusquera*.
- Se mantuvieron los diminutivos solo si designaban realidades diferentes. Así, *defectillo* se unifica con *defecto*, pero se mantiene *cuadernillo* (y no se unifica con *cuaderno*). En el caso de *chiquitito*, este diminutivo se mantiene porque *chiquito* no aparecía en los resúmenes.
- Con respecto a los números, se mantuvieron únicamente cuando indicaban años (se eliminan números de páginas, por ejemplo). Lo mismo ocurrió con los números romanos, que permanecen en el corpus siempre que se refieren a siglos (se eliminan los que se refieren a centenarios, signaturas, etc.) y se marcan de la siguiente manera: *(siglo)_XIV*.

- En los casos en los que se tuvo que hacer alguna consulta en relación con la unificación ortográfica, nos dirigimos a las siguientes obras de la Real Academia Española: el *Diccionario de la Lengua Española* (23.a ed.), el *Diccionario panhispánico de dudas* y la *Ortografía de la lengua española*.

Una vez estuvieron incorporados todos estos cambios en la lista de palabras únicas, se procedió de la siguiente manera: el corpus en formato Excel se pasó a formato '.txt' y se trabajó con el programa *BBEdit* (Figura 4), un editor de texto y HTML desarrollado por *Bare Bones Software* que ofrece una gran cantidad de funciones para la edición, búsqueda y manipulación de los corpus y datos textuales. Gracias a este programa se fueron reemplazando de forma no automática los términos originales por las propuestas de modificación del listado de palabras únicas.

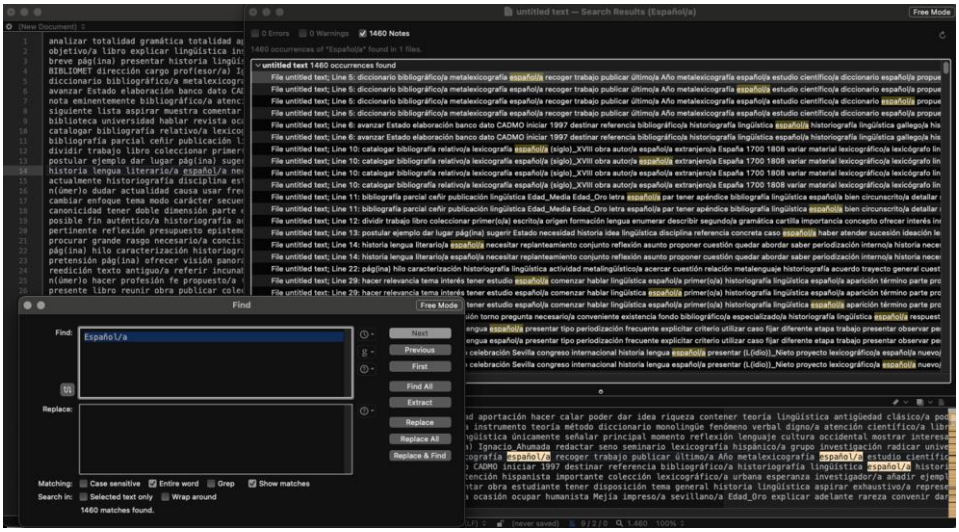


Figura 4. Visión del programa BBEdit, Fuente: elaboración propia.

Todo este proceso (Figura 5), se realizó en tres ocasiones para depurar el corpus y asegurarnos de que se siguieran de manera rigurosa los criterios explicados arriba.

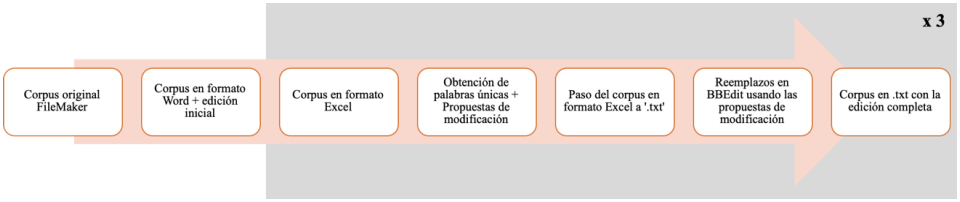


Figura 5. Pasos en el proceso de edición. Fuente: elaboración propia.

En la Tabla 1 incluimos un ejemplo del texto inicial y del texto editado tal y como queda en el corpus definitivo:

Resumen original	**Resumen editado final**
3994 "Desde 1780 no se ha efectuado una revisión sistemática y profunda del «Diccionario» de la Academia. Una revisión que reúna estas condiciones es imprescindible en nuestro tiempo, pero para ello se requiere disponer de una amplia base documental. La Academia dispone de esa documentación, en bruto, almacenada en sus ficheros léxicos. Y cabalmente, la elaboración, ordenación y transformación en un producto útil de esos materiales es la tarea de la redacción del «Diccionario histórico de la lengua española»."	1780 efectuar revisión sistemático/a profundo/a diccionario academia revisión reunir condición imprescindible tiempo requerir disponer amplio/a base documental academia documentación bruto almacenar fichero léxico/a cabalmente elaboración ordenación transformación producto útil material tarea redacción diccionario histórico/a lengua español/a

Tabla 1. Corpus original frente a Corpus editado. Fuente: elaboración propia.

Este corpus definitivo, denominado BiTe_Corpus, está publicado en abierto y disponible en Zenodo. Se trata de un archivo de texto (txt.) que finalmente contiene 102 613 palabras y 9270 palabras únicas y está especialmente concebido para el estudio metahistoriográfico de la historia de la lingüística hispánica, como veremos en los siguientes apartados.

1.2 Software empleado: Excel, Voyant, Tableau y Gephi

Para esta investigación y como herramienta auxiliar principal, hemos utilizado el programa *Microsoft Excel*, de la *suite* de *Office*. Este programa permite un manejo

sencillo de archivos csv., por ejemplo, y al ser un programa más popular que otros para el empleo de hojas de cálculo —y aunque es software propietario; esto es, es necesaria una licencia de pago, individual o institucional—, la mayor parte de *software* de análisis de datos permite importar archivos en formato xls. Este programa ofrece por su parte análisis de datos y generación de gráficos, pero es algo "ciego" en el sentido de que los ofrece por defecto y de manera muy uniforme. Puede ser, sin embargo, una forma rápida de obtener gráficos interesantes en un momento dado.

Los tres programas de análisis que hemos empleado para para analizar el BiTe_Corpus y realizar los análisis metahistoriográficos *per se* son *Voyant* (<https://voyant-tools.org/>), *Gephi* (<www.gephi.org>) y *Tableau* (<www.tableau.com>). Los dos primeros son software libre y pueden utilizarse en línea (el primero) o descargarse (el segundo). En nuestra opinión, debemos tender al empleo de programas de acceso libre porque posee numerosas ventajas frente al propietario, pero no es este el lugar para su discusión. Interesa, sobre todo, que suelen tener una comunidad de usuarios muy activa que permite la discusión, la implementación de mejoras o su mantenimiento de una forma más continua. *Tableau*, en cambio, es un software propietario de visualización de datos interactivos, pero tiene muchas opciones gratuitas de uso y, en todo caso, puede sustituirse por tantos otros.

Voyant Tools se define como "un entorno de lectura y análisis basado en web para textos digitales" y fue puesto en marcha por Stéfan Sinclair y Geoffrey Rockwell en 2016; a ambos debemos no solo esta herramienta sino la excelente monografía *Hermeneutica: Computer-Assisted Interpretation in the Humanities* también publicada en 2016. El trabajo de ambos investigadores se centra en el análisis de texto y en la implementación de técnicas digitales para el sostenimiento, además, de la herencia cultural. Como explicaremos en el apartado siguiente, *Voyant* ofrece numerosas herramientas de análisis de texto y es uno de los programas de análisis de texto más empleados en la actualidad.

Voyant es, en definitiva, una herramienta de análisis de texto en línea que permite analizar un texto, o un conjunto de textos, de una forma muy intuitiva y visual. En realidad, es especialmente útil para el análisis de textos en lenguas romances (y en otras lenguas), como el español, porque no es un programa de procedencia anglosajona —a diferencia de otros que pueden funcionar muy bien en inglés, pero no tan bien en otras lenguas—. *Voyant* también permite analizar el texto de una forma muy visual, lo que facilita la identificación de las palabras más importantes en un texto o los temas más relevantes.

Gephi es, por su parte, un programa específico de visualización y análisis de grafos, entendidos como la expresión de una red más o menos compleja, y permite al usuario identificar las relaciones entre los datos. Se puede encontrar una descripción del proyecto en Bastian *et al.* (2009), pero toda la información sobre el

programa está en su página web. Para conocer el alcance o las posibilidades de *Gephi* en las Humanidades, recomendamos la consulta y la lectura de los interesantes trabajos de Martin Grandjean (<http://www.martingrandjean.ch/>).

2. Análisis del corpus: primeros resultados metahistoriográficos

A lo largo de este epígrafe mostraremos algunas de las posibilidades que ofrece el programa *Voyant* para explorar cuantitativa y cualitativamente el corpus. Este entorno web dispone de un amplio listado de herramientas[3]; de todas ellas, solo emplearemos de momento aquellas que nos pueden ofrecer datos interesantes en relación con el corpus estudiado o, además, datos que visualizaremos posteriormente con *Gephi*. Antes de ello, no obstante, es necesario avisar de que para trabajar con algunas de las herramientas de *Voyant* se tuvo que adaptar el corpus eliminando los paréntesis y los signos auxiliares utilizados (la barra, el guion bajo y los corchetes) para que *Voyant* hiciese correctamente la separación de las unidades léxicas.

2.1 *Cirrus*, nube de palabras

Gracias a la nube de palabras se pueden explorar las palabras de alta frecuencia. Según explican Sinclair y Rockwell (2016), el color de las palabras y su posición (vertical u horizontal) en el gráfico no es significativo; sin embargo, el tamaño y su localización en el centro del gráfico sí nos da información sobre la frecuencia del término. El valor máximo de la nube de palabras que crea *Voyant* es de 500 términos (Figura 6) y el mínimo de 25 términos (Figura 7). El resultado es una imagen visual y dinámica que orienta sobre el contenido del corpus analizado.

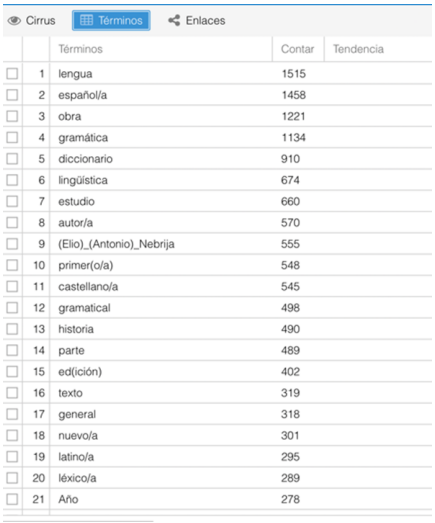
[3] Las herramientas que ofrece *Voyant* son *Bubblelines, Bubbles, Cirrus, Collocates Graph, Corpus Collocates, Contexts, Correlations, Document Terms, Corpus Terms, Documents, Knots, Mandala, MicroSearch, Phrases, Reader, ScatterPlot, StreamGraph, Summary, Terms Radio, TextualArc, Topics, Trends, Veliza, Word Tree.* Puede encontrarse una explicación sobre cada una de ellas en <https://voyant-tools.org/docs/#!/guide/about>.

Figura 6. Nube de palabras con 500 términos. Fuente: elaboración propia.

Figura 7. Nube de palabras con 25 términos. Fuente: elaboración propia.

2.2 Términos

La opción "términos" ofrece una relación de las palabras más frecuentes, pero, en este caso, muestra las palabras ordenadas por frecuencia en orden descendente. Así, de forma clara, la herramienta nos da a conocer las diez palabras más repetidas en el corpus: *lengua*, *española*, *obra*, *gramática*, *diccionario*, *lingüística*, *estudio*, *autor*/a, *(Elio)_(Antonio)_Nebrija*, *primero/a*. Esta herramienta puede servir, de manera auxiliar, para comprobar si hay errores en el corpus o para editar la lista de *stopwords*.



Cirrus | Términos | Enlaces

	Términos	Contar	Tendencia
1	lengua	1515	
2	español/a	1458	
3	obra	1221	
4	gramática	1134	
5	diccionario	910	
6	lingüística	674	
7	estudio	660	
8	autor/a	570	
9	(Elio)_(Antonio)_Nebrija	555	
10	primer(o/a)	548	
11	castellano/a	545	
12	gramatical	498	
13	historia	490	
14	parte	489	
15	ed(ición)	402	
16	texto	319	
17	general	318	
18	nuevo/a	301	
19	latino/a	295	
20	léxico/a	289	
21	Año	278	

Figura 8. Términos más frecuentes del corpus extraídos con *Voyant*. Fuente: elaboración propia.

El análisis de la frecuencia de palabras puede dar resultados interesantes si combinamos o analizamos estos datos con otro tipo de *software*. Este es el caso del siguiente ejemplo en forma de gráfico —más clásico—, para el que hemos empleado *Voyant* (frecuencia de términos), *Excel* y *Tableau*. Para este análisis, hemos buscado los años más frecuentes del corpus y hemos podido comprobar que, al menos, los diez años más frecuentes son realmente las fechas de publicación de determinadas obras (en la leyenda del Gráfico 1 se ha anotado el número de registros). De esta forma, vemos que, con la combinación de diferentes herramientas (datos en *Excel* extraídos de *Voyant*, importados a *Tableau*) y con cierto conocimiento del área (qué pueden significar determinados años), podemos obtener resultados que pueden venir a complementar o a fundamentar, llegado el caso, determinadas hipótesis en nuestras investigaciones. Y también nos permite demostrar que la *Gramática castellana* de Antonio de Nebrija es la obra más citada en el corpus de manera incontestable.

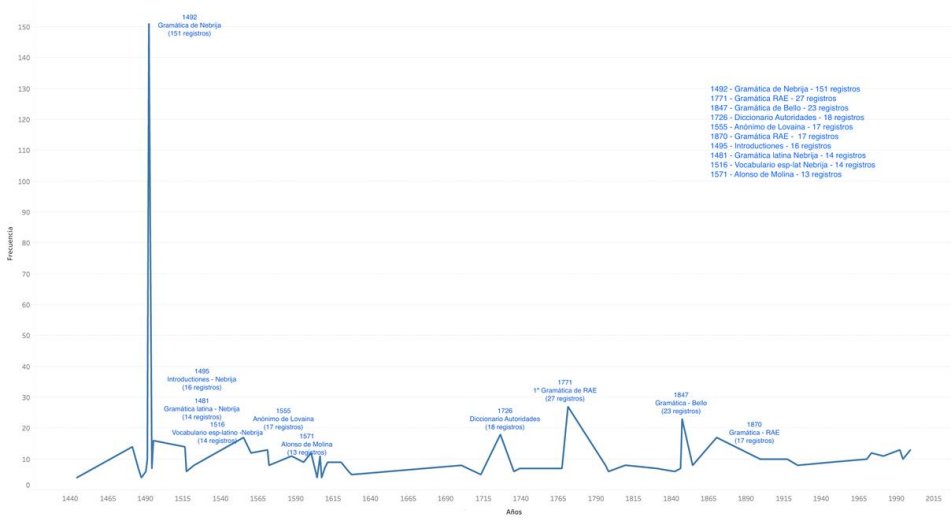


Gráfico 1. Relación de años más frecuentes = fecha de publicación. Fuente: elaboración propia.

2.3 Líneas de burbujas

La opción "líneas de burbujas" (*Bubbleline*) visualiza la frecuencia de las palabras del corpus, como hemos visto hasta ahora, pero añade una información interesante porque señala la distribución de los términos a lo largo del corpus. Esta herramienta es muy reveladora si se presenta un corpus ordenado cronológicamente o un conjunto de textos ordenados de tal forma.

Por ejemplo (Figura 9), hemos realizado una búsqueda de autores destacados de los siglos XVI y XVII (el Brocense, Juan Caramuel, Jiménez Patón, Correas, Juan de Villar, Covarrubias y Francisco del Rosal) y, como se puede apreciar, esta herramienta no solo nos da información sobre quiénes fueron los autores más citados en el corpus (destacan El Brocense, Covarrubias y Correas), sino que también evidencia su estructura. Como se explicó más arriba, no se modificó el orden de aparición de las palabras dentro de cada uno de los resúmenes y se mantuvo el orden de cada resumen dentro de los quince capítulos de BiTe, que tiene una disposición de capítulos mayormente cronológica. Esta es la razón por la que en la Figura 9 las burbujas se concentran hacia la mitad del corpus, que se correspondería con los capítulos IX y X de BiTe, "Gramática y ortografía en España en los siglos XVI y XVII" y "El nacimiento de la lexicografía monolingüe española", los dos últimos capítulos del primer tomo.

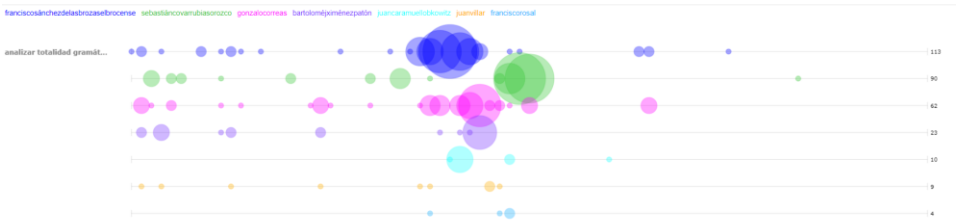


Figura 9. Líneas de burbujas con autores de los siglos XVI y XVII. Fuente: elaboración propia.

Si en la búsqueda seleccionamos autores del siglo XIX (por ejemplo, Real Academia Española, Gómez Hermosilla, Salvá, Bello o Benot), las burbujas se concentran en la parte final (i.e. a la derecha de la imagen) del corpus (Figura 10), que se corresponde con el capítulo XIV de BiTe, titulado "La lingüística en el ámbito hispanohablante: siglo XIX", aunque también se advierte que estos autores están presentes en su comienzo, concretamente en el capítulo relativo a "Materiales".

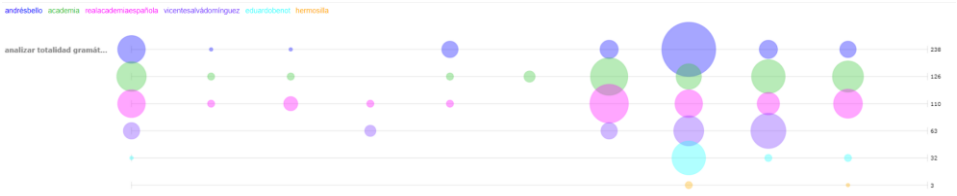


Figura 10. Líneas de burbujas con autores del siglo XIX. Fuente: elaboración propia.

2.4 *Loom*

La Figura 11 muestra un "telar" de todo el corpus y también puede ayudar a comprender cómo se distribuyen las palabras en relación con el resto (y a la vez). En las tres siguientes figuras, se puede apreciar cómo la palabra "español/a" se mantiene a lo largo de todo el corpus (Figura 11), cómo "Nebrija" aparece menos frecuentemente en la segunda parte (Figura 12) y cómo "Bello", en cambio, aumenta en la parte final (Figura 13); todo ello, en clara consonancia con los capítulos de BiTe como ocurre con algunas de las herramientas anteriores.

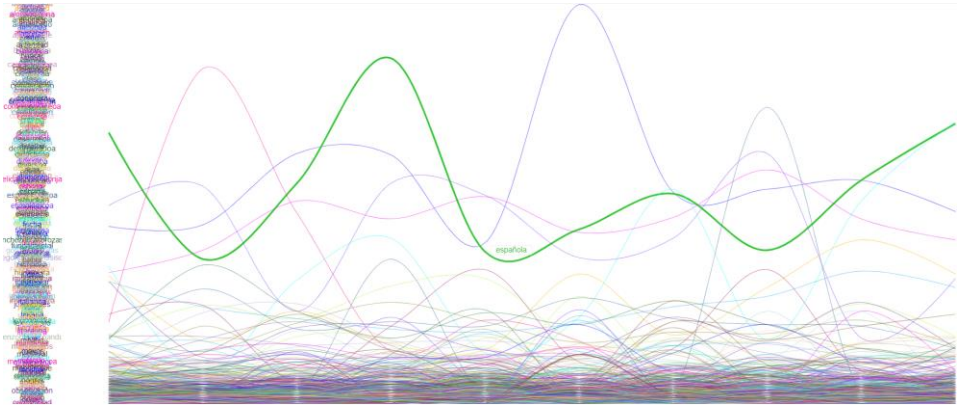


Figura 11. *Loom* con todo el corpus en el que se resalta la palabra "español/a".
Fuente: elaboración propia.

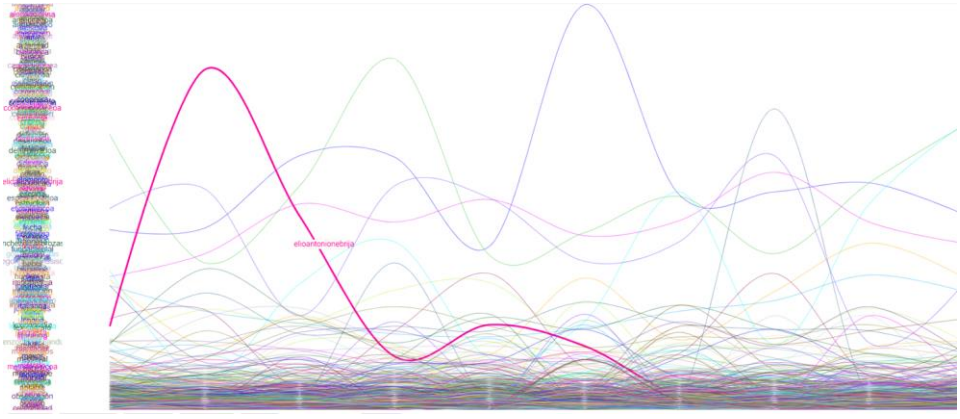


Figura 12. Loom con todo el corpus en el que se destaca "Antonio de Nebrija".
Fuente: elaboración propia.

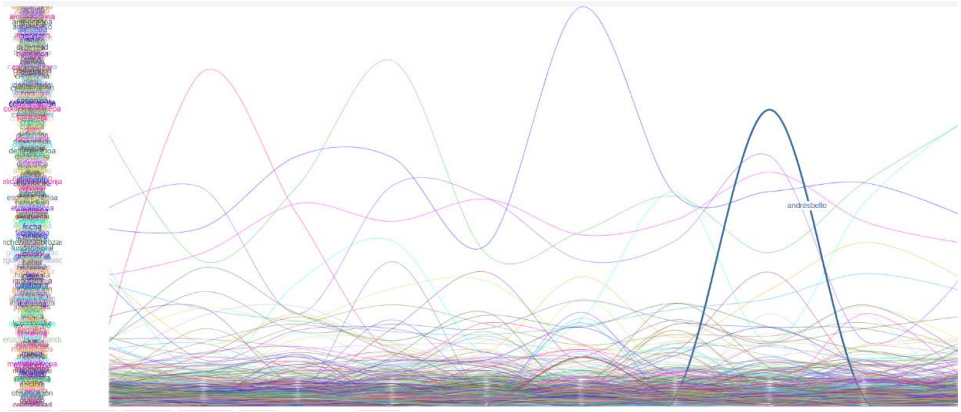


Figura 13. Loom con todo el corpus en el que se destaca "Andrés Bello".
Fuente: elaboración propia.

2.5 Búsqueda micro

En la opción de "Búsqueda micro", el corpus se representa como un bloque vertical en color gris y, sobre este bloque, se señalan en rojo los términos buscados (de diferente intensidad según la frecuencia). En las Tablas 2 y 3 se muestra, por un lado, el resultado de las búsquedas relacionadas con los autores de los primeros diccionarios bilingües, lexicografía plurilingüe y los inicios de la lexicografía monolingüe; y, por otro, las búsquedas relacionadas con las obras de la Real Academia Española, la lexicografía no académica y otros autores de los siglos XIX y XX (no necesariamente lexicógrafos). De nuevo, las correspondencias con los capítulos de BiTe son manifiestas puesto que las investigaciones sobre el nacimiento de los estudios sobre español ocupan la primera parte del corpus, mientras que los estudios centrados en autores a partir del siglo XVII se sitúan de la mitad hacia el final del corpus.

Fernández de Palencia, Nebrija, Fernández de Santaella, Pedro de Alcalá	Calepino, Molina, Gilberti, Oudin, Las Casas, Percival	Del Rosal, Covarrubias, Correas
		

Tabla 2. Búsqueda micro en *Voyant* (I). Fuente: elaboración propia.

Obras de la Real Academia Española	Terreros y Pando, Salvá, Cuervo	Bello, Hermosilla, Benot
		

Tabla 3. Búsqueda micro en *Voyant* (II). Fuente: elaboración propia.

2.6 Gráfico de flujo

El "gráfico de flujo" representa el cambio de la frecuencia de las palabras dentro del corpus, por lo que es una forma muy eficaz de comparar el comportamiento de determinados términos. Así, en la Figura 14 se aprecia que, si hacemos una cala con autores destacados de los siglos XVI y XVII, su aparición en el corpus se limita a la primera parte (tomo 1 de BiTe, capítulo VI "Ideas, teorías y polémicas sobre el lenguaje y la lengua en el Siglo de Oro", capítulo IX "Gramática y ortografía en España en los siglos XVI y XVII" y capítulo X "El nacimiento de la lexicografía monolingüe española"); mientras que, como muestra la Figura 15, si se buscan autores de los siglos XIX y XX, cambia de forma clara el lugar que estos últimos ocupan en el corpus (hacia el final corpus, tomo 2 de BiTe, capítulo XIV "La lingüística en el ámbito hispanohablante: siglo XIX" y Capítulo XV "Materiales para una historia de la lingüística en el ámbito hispanohablante en el siglo XX").

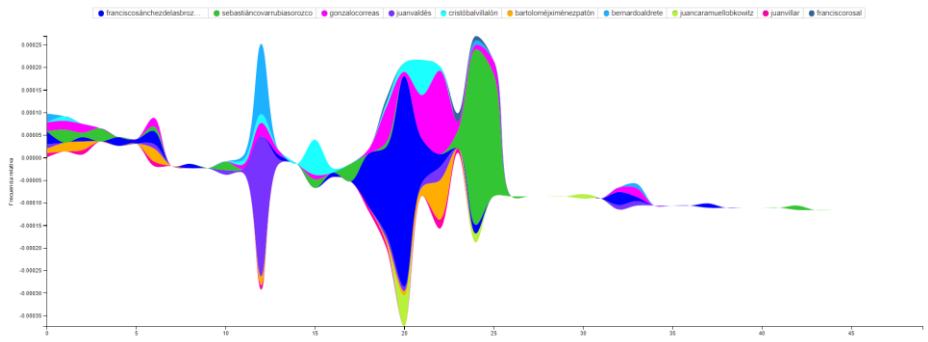


Figura 14. Gráfico de flujo autores de los siglos XVI y XVII. Fuente: elaboración propia.

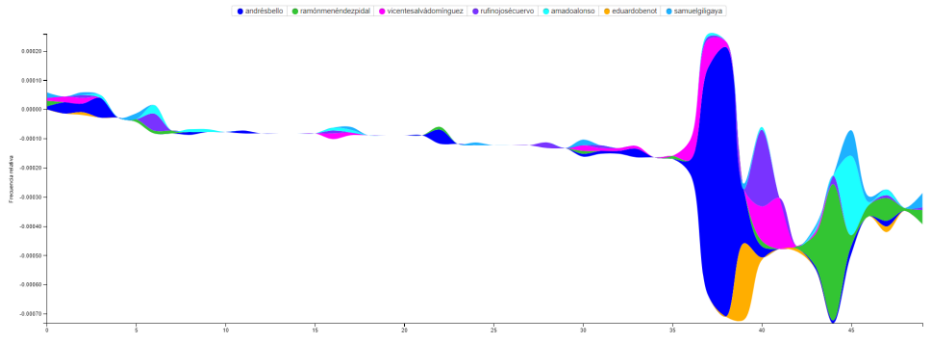


Figura 15. Gráfico de flujo con autores de los siglos XIX y XX. Fuente: elaboración propia.

El análisis de flujo de la Figura 16 compara el uso del término "castellano/a" y el de "español/a" (es decir, su frecuencia de aparición) y su distribución a lo largo del corpus. Incluimos este ejemplo porque más adelante analizaremos algunos otros aspectos del comportamiento de estos términos.

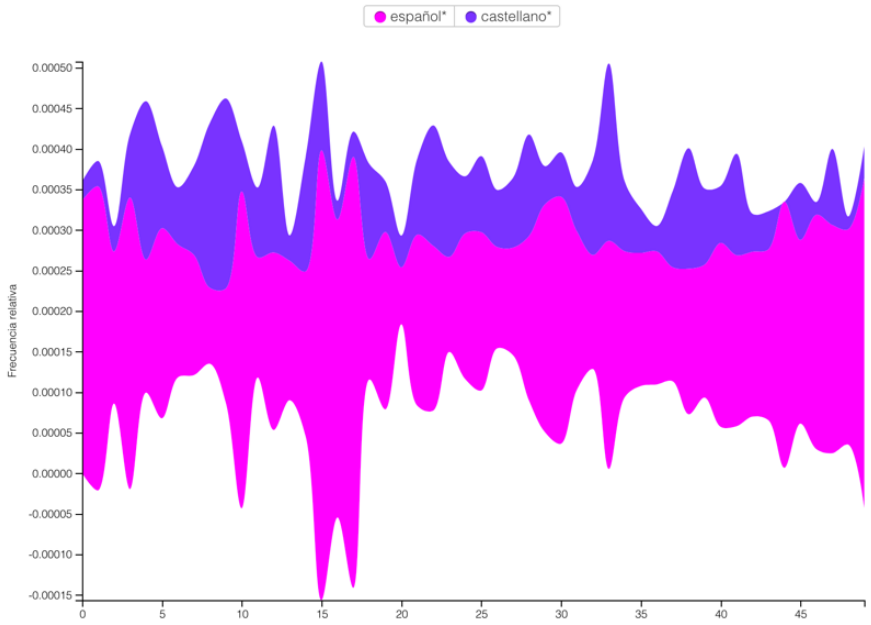


Figura 16. Gráfico de flujo para los términos "castellano" y "español". Fuente: elaboración propia.

2.7 Tendencias

La opción "tendencia", parecida al gráfico de flujo, muestra en un gráfico la frecuencia del término para cada segmento en el corpus; estos segmentos se crean automáticamente y son todos aproximadamente de la misma longitud (Sinclair y Rockwell, 2016). En las siguientes figuras, se puede observar cómo se distribuyen los cinco autores más citados a lo largo del corpus (Figura 17) y lo que ocurre concretamente, por ejemplo, con María Moliner (Figura 18), presente en el último capítulo de BiTe.

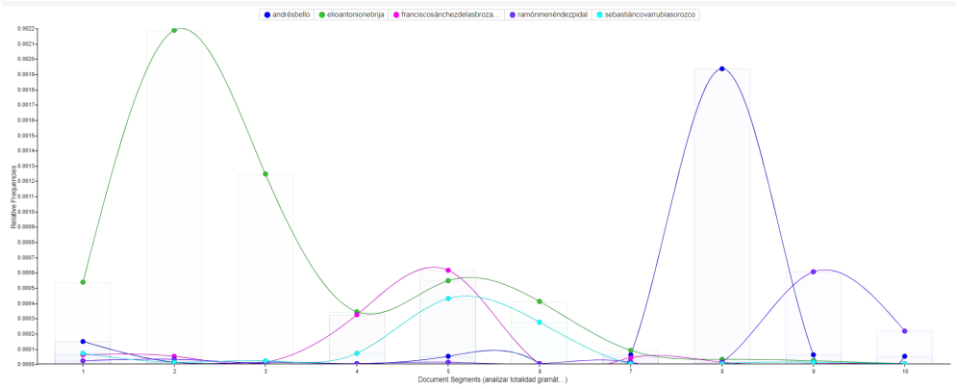


Figura 17. Tendencias con los cinco autores más citados. Fuente: elaboración propia.

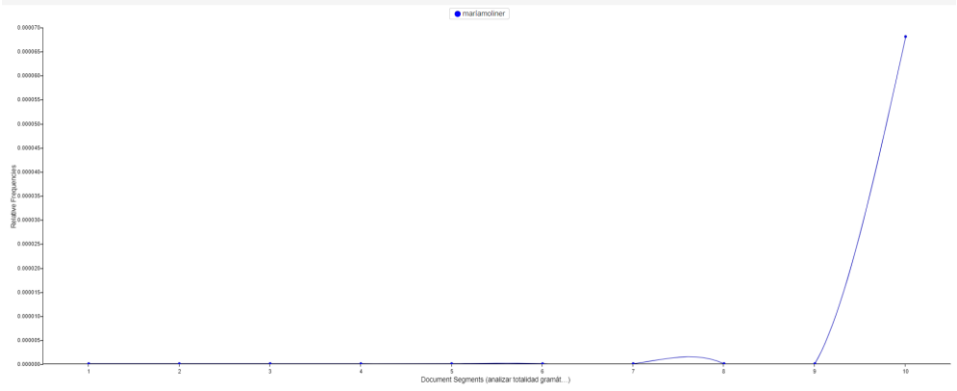


Figura 18. Tendencias con María Moliner. Fuente: elaboración propia.

2.8 Árbol de palabras

Por último, y aunque *Voyant* ofrece muchas más posibilidades de análisis —por razones de espacio no vamos a incluir ejemplos de cada una de ellas—, hemos querido emplear el denominado Árbol de palabras de un término (Figura 19). Ahora bien, el árbol se construye según un número limitado de concordancias para la palabra clave y las ramas que se muestran no necesariamente se basan en la frecuencia (Sinclair y Rockwell, 2016). Por ello, es más interesante trabajar con otros programas (por ejemplo, *Gephi*, como veremos en la siguiente sección) para obtener una visualización precisa de las relaciones (y su importancia) entre los términos del corpus.

Figura 19. Árboles de palabras de Bello, Moliner y Nebrija. Fuente: elaboración propia.

Como conclusión parcial del empleo de *Voyant*, consideramos que este recurso presenta varias ventajas: no es necesario descargar ninguna aplicación porque se puede trabajar directamente en línea, permite manejar sus herramientas de modo bastante sencillo e intuitivo y está preparada para poder compartir los resultados a través de la web, incrustando los gráficos, tablas, etc. Además, aunque en este caso hemos utilizado un único corpus, *Voyant* es especialmente interesante para el cotejo de varias obras dentro de un mismo corpus.

3. Exploración del corpus mediante visualización de redes con *Gephi*

En general, los estudios lingüísticos que trabajan de manera cuantitativa con las palabras suelen centrarse en la frecuencia; esto es, en el número de apariciones de una palabra determinada en un corpus. Se trata, por decirlo en cierto modo, de un estudio esencialmente "paradigmático". No obstante, gracias al desarrollo tecnológico —entro otros—, la lingüística de corpus (McEnery y Hardie 2011) ofrece hoy en día nuevas formas de realizar estudios estilométricos y está avanzando en la precisión y en la profundidad de los análisis textuales mediante diferentes herramientas y métodos. Una de estas nuevas formas de acercarse al análisis lingüístico lo ofrece la Teoría de redes complejas matemáticas (Albert y Barabasi, 2002; Criado *et al.* 2011).

Partiendo de la base de que una lengua —por sus propiedades— puede tratarse como una red compleja (Cong *et al.*, 2014; Martinčić-Ipšić *et al.*, 2016), cada vez son más numerosas las investigaciones lingüísticas de todo tipo que se realizan desde esta perspectiva (v., por ejemplo, Chen 2014). En esta línea, Köhler (2014) compara cómo se tratan los datos lingüísticos extraídos a partir de la investigación de las redes lingüísticas y los datos extraídos a partir de la lingüística cuantitativa tradicional, poniendo de manifiesto las ventajas asociadas a la utilización del primer tipo de tratamiento de datos. Criado-Alonso *et al.* (2020, 2021), por ejemplo, han extraído las propiedades lingüísticas de un corpus de textos a

partir de la teoría de redes complejas con idea de describir las características estilísticas y topológicas del lenguaje de especialidad —en su caso, el lenguaje de especialidad matemático—.

Siguiendo este ejemplo, vamos a exponer algunos resultados metahistoriográficos derivados del estudio de nuestro corpus a través de la visualización de grafos. Para ello nos vamos a centrar en el estudio de las denominadas "colocaciones" (Firth 1957, Halliday y Mattthiesen 2004) del lenguaje de especialidad —en nuestro caso, de la historiografía lingüística hispánica—o en la relación entre determinados conceptos o palabras a través de diferentes algoritmos expresados en forma de grafos específicos del programa *Gephi*. En definitiva, veremos que el análisis mediante la visualización de redes puede ser esencialmente significativo para el estudio de las relaciones sintagmáticas —a diferencia de los estudios cuantitativos más tradicionales— y, por extensión, pueden ofrecer nuevas perspectivas de trabajo y análisis metahistoriográficos.

La herramienta de Voyant denominada "Colocaciones" permite extraer, para una palabra seleccionada (por ejemplo, "gramática"), la relación de palabras que la acompañan y el número de veces que se da esta combinación: por ejemplo, "gramática" aparece junto a la palabra "castellana" un total de 186 veces en nuestro corpus. De esta forma, podemos contar con tres datos (que vamos a denominar *source, target* y *weight*) que suponen un nodo de origen ("gramática"), un nodo de llegada ("castellana") y una cifra (186) como peso de la arista (la expresión de la cantidad de veces que aparece esta combinación de palabras) que nos va a permitir construir una red y visualizar la información en forma de grafo. Exportada esta lista de *Voyant* a Excel, podemos comenzar a trabajar entonces con programas de visualización como *Gephi* o *Tableau*.

En un principio, estas herramientas han sido empleadas con intención de realizar un estudio estilométrico. Como es lógico, solo la herramienta en sí ya ofrece resultados que, en un momento dado, pueden analizarse desde un punto de vista meramente cuantitativo. Sin embargo, el conocimiento de la disciplina es fundamental para poder extraer conclusiones relevantes o, incluso, para realizar según qué búsquedas. Como podemos suponer, este tipo de estudio puede ofrecer datos complementarios muy interesantes al análisis historiográfico tradicional; no obstante, en este punto defendemos —como veremos— que también puede funcionar como una potente herramienta metahistoriográfica, lo que realmente nos permite sostener el objetivo de este trabajo: a saber, que el análisis de datos es un medio de investigación —no un fin en sí mismo— y que este funciona realmente cuando es realizado por especialistas en la materia formados asimismo en la teoría o teorías de análisis de datos.

Veamos algunos ejemplos a propósito del uso o comportamiento de los términos "español" y "castellano" en la bibliografía secundaria de historia de la lin-

güística española[4]. En primer lugar, nos detendremos en aspectos más cuantitativos —para ello utilizaremos *Tableau*; Figura 20—; y, en segundo lugar, estudiaremos aspectos algo más cualitativos mediante los grafos obtenidos con *Gephi* con los mismos datos.

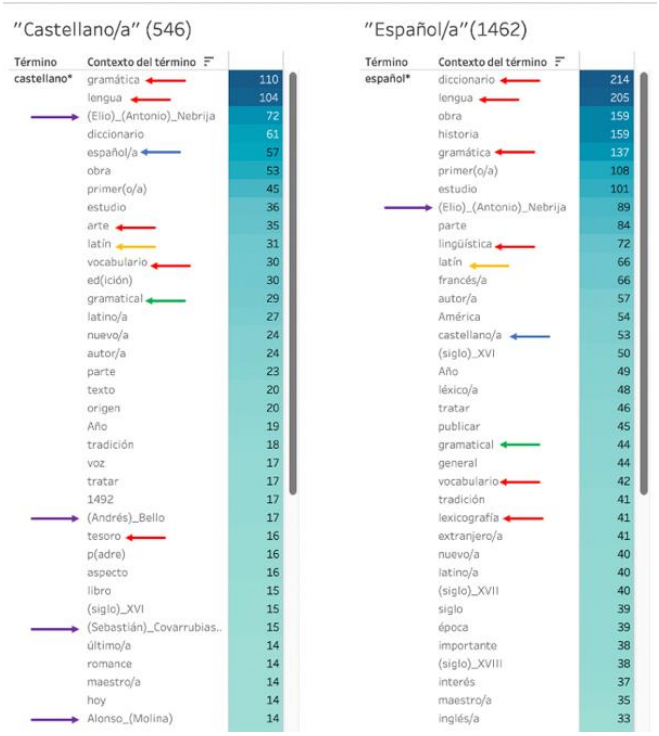


Figura 20. Colocaciones más frecuentes de "castellano/a" y "español/a" en BiTe_Corpus. Las flechas han sido incluidas por las autoras para destacar determinados términos.
Fuente: elaboración propia.

El estudio de las colocaciones para "castellano/a" y "español/a" y el número de apariciones permite apreciar, en un primer momento, usos muy claros (y muy determinados) referidos a uno y otro término cuando, en determinados contextos, podrían parecer sinónimos. Lo primero que llama la atención es, sin dudarlo, la mayor frecuencia del término "español/a" (1462 apariciones) frente a "castellano/a", que solo cuenta con 546.

4 En este apartado, estos términos serán empleados de manera genérica (es decir, que ambos incluyen sus respectivas variantes de género y número). Hemos de tener en cuenta también que el uso de "español" o "castellano" en un momento dado puede deberse, como veremos, a la palabra del título de una obra concreta, a un modo de denominar la lengua, etcétera.

En la Figura 20 podemos ver una imagen extraída de *Tableau* con los datos obtenidos previamente por *Voyant* y *Excel*. Así, la columna de la izquierda relaciona los términos que aparecen con "castellano/a" y que suponen un total de 546 distribuidos como aparecen en la columna "de calor" (más o menos fuerte) de color azul. El término más habitual que aparece combinado con "castellano/a" es "gramática" —con 110 apariciones— seguido de "lengua", que aparece 104 veces. En el corpus, por tanto, "gramática castellana" aparece un total de 110 veces y "lengua castellana", 104. Llama la atención que el tercer término más habitual en combinación con "castellano/a" sea un autor; sin embargo, el hecho de que sea el mismo Nebrija explicaría el comportamiento de este término con "castellano/a", dado el título de la gramática que publicó en 1492 —año que también aparece en la relación de palabras—.

En lo que se refiere al término "español/a", observamos que la combinación de palabras es algo desigual comparada con la de "castellano/a". Así, además de ser un término que aparece con mucha más frecuencia, salvando la combinación "lengua española" —que, creemos, tiene también su explicación por las veces que puede aparecer en el título de distintas obras, como en el caso de la *Gramática castellana* de Nebrija—, aparece con palabras diferentes: "diccionario", "obra" e "histórica" son los términos más comunes; el primer autor citado es también Nebrija, pero en una posición más baja que en el caso de "castellano/a"; y, entre otras apreciaciones más que se pueden realizar, ciertamente este término aparece con otros tantos relacionados con la lexicografía. Si nos permiten esta explicación, podríamos decir que "castellano/a" se refiere más al "arte" y "español/a" a "lingüística": es decir, que hay cierto barniz histórico en la misma elección del término.

En esta relación, llama asimismo la atención la presencia de determinados autores asociados a "castellano/a" y a "español/a" que son realmente muy diferentes. Como hemos visto, en ambos casos el autor más frecuente es Nebrija —que, por otra parte, es el autor (y su obra, la *Gramática*) que más apariciones presenta en el corpus—; pero, quitando este caso, la nómina de autores es muy desigual. En el caso de "castellano/a", encontramos también a Andrés Bello, a Covarrubias y a Alonso de Molina en estos primeros términos. Pero en el caso de la palabra "español/a", y solo en posiciones muy inferiores —tanto que no aparece en la imagen—, el autor que más habitualmente aparece junto a dicho término es Ramón Menéndez Pidal.

Aunque de manera algo más sutil, la diferencia de uso entre estos términos puede observarse en la representación, en este caso, de las relaciones entre los términos (además de su frecuencia). Esta representación de una red compleja es lo que denominamos "grafos" y lo que puede proporcionar información complementaria pero relevante a los análisis cualitativos y/o cuantitativos más "tradicionales" que hemos visto en apartados anteriores.

Siguiendo con el ejemplo de "castellano/a" y "español/a", es interesante observar que para el primero hay aristas proporcionalmente más importantes (con más peso) que en el caso de "español/a" —aunque hayamos visto que las palabras y colocaciones para este término tengan más frecuencia— (v. Figura 21). Esto quiere decir que se utilizan con más consistencia —y más veces, proporcionalmente— las combinaciones de "castellano/a con "gramática", "lengua", "Nebrija" e —incluso— "diccionario", que las de "diccionario", "lengua" u "obra" en el caso de "español/a": es tan sencillo como fijarse en que el grosor de las aristas de las figuras 22, 23 y 24 es mayor que el grosor de las aristas de la imagen derecha.

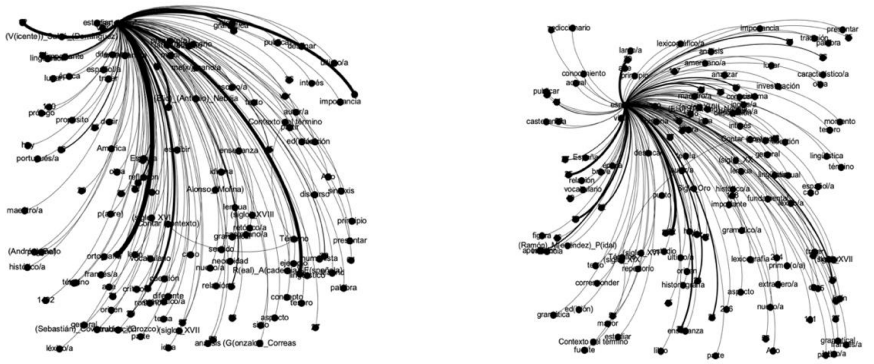


Figura 21. Grafos de colocaciones más habituales de los términos "castellano/a" (imagen izquierda) y "español/a" (imagen derecha). Fuente: Elaboración propia.

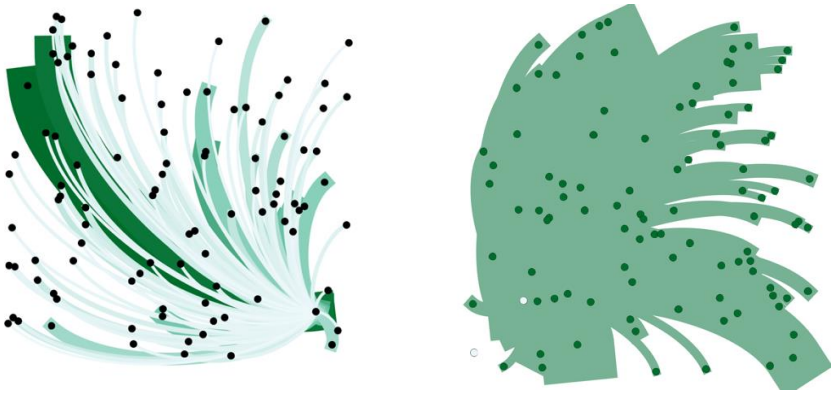


Figura 22. Grafos de colocaciones más habituales de los términos "castellano/a" (imagen izquierda) y "español/a" (imagen derecha). Las diferencias de color y grosor expresan más o menos "desproporción". Fuente: Elaboración propia.

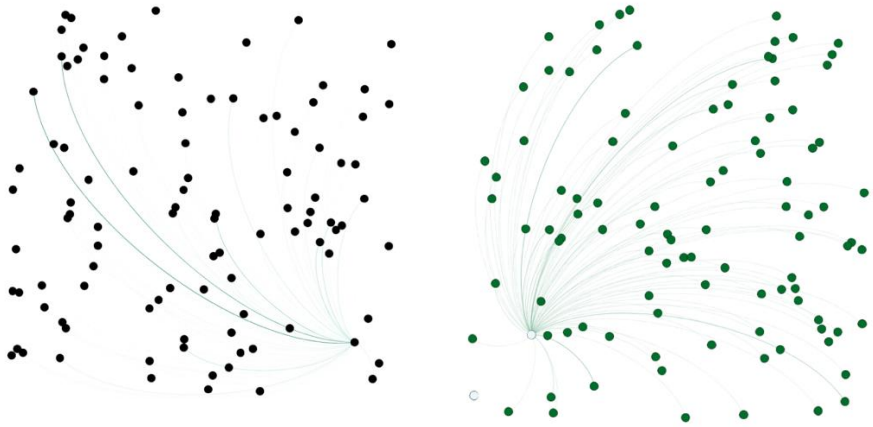


Figura 23. Grafos de colocaciones más habituales de los términos "castellano/a" (imagen izquierda) y "español/a" (imagen derecha). Las diferencias de grosor expresan más o menos "desproporción". Fuente: elaboración propia.

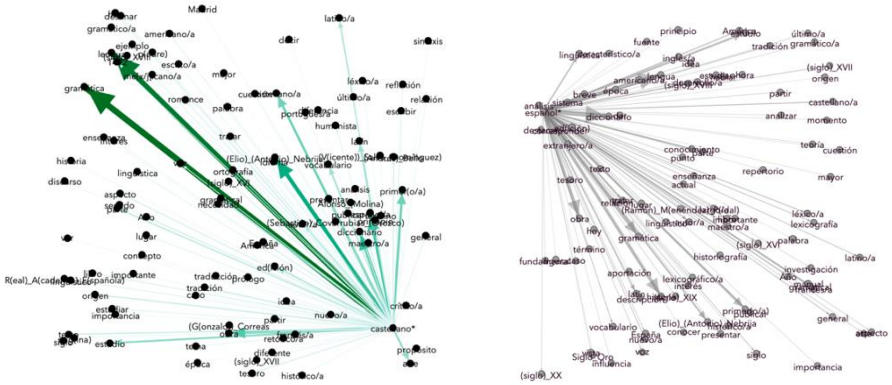


Figura 24. Grafos de colocaciones más habituales de los términos "castellano/a" (imagen izquierda) y "español/a" (imagen derecha). Las diferencias de grosor y color de las aristas expresan más o menos "desproporción". Fuente: elaboración propia.

En este punto, es necesario señalar que *Gephi* permite visualizar la misma red con diferentes algoritmos, de forma que incluso la información puede mostrarse de otras maneras que pueden ser también reveladoras. Aunque es posible que no se

aprecie el detalle, este otro algoritmo de *Gephi* (el algoritmo *Yifan-Hu*) nos muestra otro tipo de relación entre los términos "castellano/a" (en azul, hacia la izquierda) y a "español/a" (en verde, hacia la derecha); en este caso, qué términos tienen en común y cuáles no, puesto que se trata de un algoritmo de repulsión (Figura 25): así, por ejemplo, para "castellano/a" encontramos (a la izquierda) a Alonso de Molina y para "español/a" a Menéndez Pidal (a la derecha). Nebrija, en cambio, aparece en el conjunto del centro.

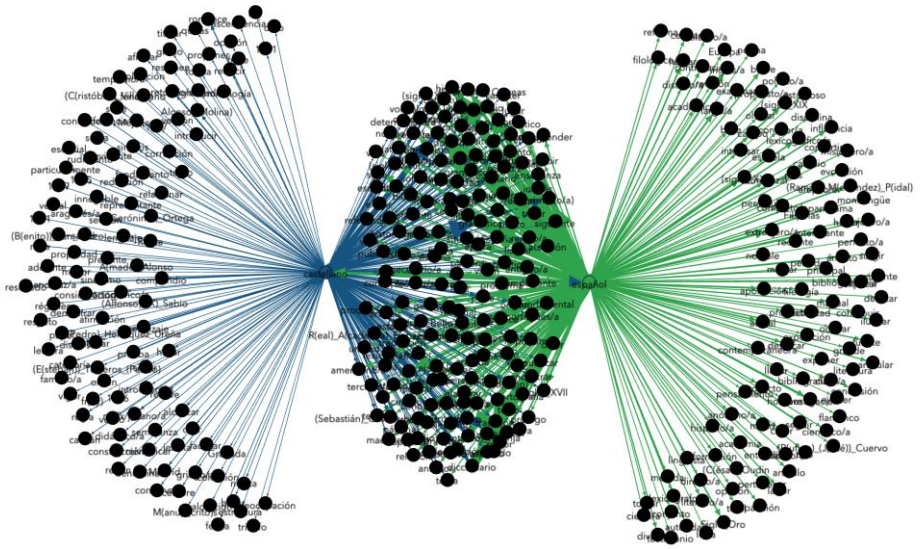


Figura 25. Grafo con algoritmo *Yifan Hu* de colocaciones más habituales de los términos "castellano/a" (izquierda) y "español/a" (derecha). Fuente: elaboración propia.

En realidad, muchísimos son los análisis, visualizaciones y conclusiones que podemos realizar y extraer con este tipo de herramientas. Como señalábamos en su descripción, con *Gephi* podemos visualizar la información desde la frecuencia, como con otros programas, pero es especialmente interesante si lo que queremos observar son las relaciones entre los términos y las conclusiones a las que, en un momento dado, podemos llegar a partir de ese análisis de redes concreto. Como hemos visto con este análisis de los términos "español/a" y "castellano/a", estas visualizaciones concebidas como "objetos de investigación" de nuestro corpus se amplían enormemente si no nos quedamos únicamente en los estudios de frecuen-

cia —que también ofrecen, no obstante, interesantísimas posibilidades para el análisis historiográfico y metahistoriográfico— y comenzamos a incorporar los análisis de redes, entre otros, a nuestra metodología de investigación.

4. Comentarios finales

Este artículo se compone de dos partes principales: la presentación y definición del denominado BiTe_Corpus y el análisis de los "objetos de estudio" obtenidos mediante las diferentes herramientas digitales —programas como *Voyant* o *Gephi*— sobre dicho corpus.

Señalábamos en el resumen de este artículo que nuestro objetivo era mostrar una forma de acercarse a la historiografía lingüística hispánica mediante ciertas herramientas digitales. Puesto que estas ofrecen lo que hemos denominado nuevos "objetos de estudio", a lo largo de los apartados de este trabajo hemos incorporado diferentes gráficos, visualizaciones o tablas que nos acercan, desde un punto de vista menos convencional en nuestra área, a lo que ha ocurrido ciertamente en la historiografía lingüística hispánica a lo largo de los años. Así, desde el BiTe_Corpus y empleando programas como *Excel*, *Voyant* o *Gephi*, obtenemos positivamente nuevas posibilidades de análisis y de investigación; en nuestro caso, y con este corpus, se trata efectivamente de un estudio metahistoriográfico, ya que entendemos como tal "el trabajo reflexivo sobre la labor historiográfica, y especialmente por lo que se refiere a sus aspectos metodológicos y teóricos" (Swiggers 2009, 71).

Así, en el caso de la historia de la lingüística hispánica, podemos emplear estas herramientas para analizar diferentes autores, obras o temas de estudio en la historiografía lingüística española y evaluar si reflejan una visión muy concreta de la historia de la lingüística o si, por el contrario, ofrecen una visión más amplia que incluye su análisis en diferentes contextos sociales y culturales. En cualquier caso, la posibilidad de poder cruzar y comparar los resultados de estudios de este tipo, y poder visualizarlos en diferentes formas, puede dar lugar a una reflexión más profunda y completa sobre la historia de la lingüística hispánica y de la forma como la historiografía lingüística española ha entendido y entiende la historia de nuestra especialidad.

La idea o el objetivo de mostrar las posibilidades de los análisis de este tipo sobre materias como la historio o la historiografía lingüística tiene que ver sobre todo con invitar a realizar investigaciones siguiendo esta metodología —o parecida; no es, ni mucho menos, *definitiva*—, teniendo en cuenta que la historiografía lingüística hispánica es, sin ninguna duda, pionera en muchos aspectos en el área internacional de la historia de la lingüística. Pero para poder llevar a cabo una investigación de este tipo, con todas las posibilidades que abre, es fundamental

contar con un corpus muy bien definido y con un conocimiento especializado de la materia en cuestión. Solo el *software* o solo la visualización de un grafo, por ejemplo, no son necesariamente reveladores por sí mismos. Pero, en cambio, correctamente empleadas como herramientas de investigación de un área de especialidad, puede mostrarnos una realidad teórica que, de otra manera, solo podíamos intuir.

Referencias bibliográficas

Albert, Reka & Barabasi, Albert-Laszlo. 2002. "Statistical mechanics of complex networks". En: *Reviews of Modern Physics* 4, 47-97.

Bare Bones Software. s. f. *BBEdit 14.5*. Disponible en <https://www.barebones.com> [Fecha de consulta: 18/07/2022].

Bastian, Mathieu & Heymann, Sebastien & Jacomy, Mathieu. 2009. "Gephi: an open-source software for exploring and manipulating networks". En: *International AAAI Conference on Weblogs and Social Media*, 361-362.

Battaner Moro, Elena. 2009. "La investigación sobre ortografía, fonética y fonología en la tradición lingüística española". En: Bastardín Candón, Teresa & Rivas Zancarrón, Manuel & García Martín, José María (eds.), *Estudios de historiografía lingüística*. Cádiz: Universidad de Cádiz, 27-44.

Battaner Moro, Elena. 2018. "Herramientas digitales e información bibliográfica especializada en la Historiografía lingüística hispánica: El Proyecto BiTe-AP1 (2008-2020)". En: Díaz Ferro, Marta *et al*. (eds.), *Actas do XIII Congreso Internacional de Lingüística Xeral: Vigo, 13-15 de xuño de 2018*. Vigo: Universidade de Vigo, 118-125.

Battaner Moro, Elena & Esparza Torres, Miguel Ángel. 2018. "La «Bibliografía Temática de Historiografía Lingüística española – Apéndice 1» (2008-2020). Proyecto historiográfico". En: *Boletín de la Sociedad Española de Historiografía Lingüística* 12, 35-52. Disponible en este enlace.

Battaner Moro, Elena & Esparza Torres, Miguel Ángel (coords). Con la colaboración de Acevedo López, Víctor & Fernández de Gobeo, Nerea & Gil de la Puerta, Macarena & Herranz Llácer, Cristina & López Iniesta, Juan Alonso & Segovia Gordillo, Ana. 2022. *Bibliografía temática de historiografía lingüística española – Apéndice 1 (2008-2020)*. Disponible en <www.biteap1.com> [Fecha de consulta: 02/06/2022].

Battaner Moro, Elena & Herranz-Llácer, Cristina V. & Segovia Gordillo, Ana. 2022. *BiTe_Corpus* (Version 01) [Data set]. Zenodo. Disponible en este enlace.

Calvo Fernández, Vicente & Esparza Torres, Miguel Ángel. 2008. "The incoporation of aspects of Textual Linguistics and Pragmatics into the Historiographical research of Spanish linguistics". En: *Beiträge Zur Geschichte Der Sprachwissenschaft* 18.2, 275-294.

Chen, Xinying. 2014. "Language as a whole – A new framework for linguistic knowledge integration: Comment on 'Approaching human language with complex networks' by Cong and Liu". En: *Physics of Life Reviews* 11.4, 628-629. Disponible en este enlace.

Cong, Jin & Liu, Haitao. 2014. "Approaching human language with complex networks". En: *Physics of life reviews* 11.4, 598-618.

Criado, Regino & Flores, Julio & García del Amo, Alejandro & Romance, Miguel. 2011. "Analytical relationships between metric and centrality measures of a network and its dual". En: *Journal of Computational and Applied Mathematics* 235.7, 1775-1780.

Criado-Alonso, Ángeles & Battaner-Moro, Elena & Aleja, David & Romance, Miguel & Criado, Regino. 2020. "Using complex networks to identify patterns in specialty mathematical language: a new approach". En: *Social Network Analysis and Mining* 10.1, 1-10. Springer.

Criado-Alonso, Ángeles & Battaner-Moro, Elena & Aleja, David & Romance, Miguel & Criado, Regino. 2021. "Enriched line graph: A new structure for searching language collocations" En: *CHAOS. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena* 142, 110509. Disponible en este enlace.

Cuartero Sánchez, Juan Manuel. 2002. "«Significado léxico» y «significado gramatical» en las gramáticas del español moderno". En: *Boletín de la Sociedad Española de Historiografía Lingüística* 3, 43-78.

Esparza Torres, Miguel Ángel. 2006. "Materiales para una historia de la lingüística española: La Bibliografía Temática de Historiografía Lingüística Española". En: Roldán Pérez, Antonio (ed.), *Caminos actuales de la historiografía lingüística: actas del V Congreso Internacional de la Sociedad Española de Historiografía lingüística*, vol. 1. Murcia: Universidad de Murcia, 517-528.

Esparza Torres, Miguel Ángel. 2007. "Los inicios de la lexicografía en España". En: Dorta Luis, Josefa & Corrales Zumbado, Cristóbal J. & Corbella Díaz, Dolores (eds.), *Historiografía de la lingüística en el ámbito hispánico*. Madrid: Arco Libros, 231-268.

Esparza Torres, Miguel Ángel. 2010. "Dimensiones de la lingüística misionera española". En: Assunçao, Carlos & Fernandes, Gonçalo & Loureiro, Marlene (eds.), *Ideias linguísticas na Península Ibérica (séc. XIV a séc. XIX): projeção da linguística ibérica na América Latina e Ásia*, vol. 1. Münster: Nodus Publikationen, 201-214.

Esparza Torres, Miguel Ángel (dir.) & Battaner Moro, Elena & Calvo Fernández, Vicente & Álvarez Fernández, Adrián & Rodríguez Barcia, Susana. 2008. *Bibliografía Temática de Historiografía Lingüística Española. Fuentes Secundarias*. Hamburg: Helmut Buske Verlag.

Fernández de Gobeo Díaz de Durana, Nerea & Gil de la Puerta, Macarena & Acevedo López, Víctor F. 2021. "Análisis cualitativo y cuantitativo de los materiales registrados en BiTe-Ap1: Gramática escolar, sintaxis y lingüística misionera". En: *Boletín de la Sociedad Española de Historiografía Lingüística* 15, 43-69. Disponible en este enlace.

Fernández Juncal, María del Carmen. 2013. *Léxico disponible en Cantabria. Estudio sociolingüístico*. Salamanca: Ediciones Universidad de Salamanca.

Firth, John R. 1957. "A Synopsis of Linguistic Theory, 1930-1955". En: *Studies in Linguistic Analysis*, Special Volume, 1-32.

Halliday, Michael A. K. & Matthiessen, Christian M. I. M. 2004. *Introduction to Functional Grammar* (Third edition). London & New York: Routledge, Taylor & Francis Group.

Köhler, Reinhard. 2014. "Linguistic complex networks as a young field of quantitative linguistics: Comment on 'Approaching human language with complex networks' by J. Cong and H. Liu". En: *Physics of Life Reviews* 11.4, 630-631.

Martinčić-Ipšić, Sandra & Margan, Domagoj & Meštrović, Ana. 2016. "Multilayer network of language: A unified framework for structural analysis of linguistic subsystems". En: *Physica A* 457, 117-128.

McEnery, Tony & Hardie, Andrew. 2011. *Corpus Linguistics: Method, Theory and Practice. (Cambridge Textbooks in Linguistics)*. Cambridge: Cambridge University Press.

Real Academia Española. 2014. *Diccionario de la lengua española* (23.ª ed.). Madrid: Espasa.

Real Academia Española & Asociación de Academias de la Lengua Española. 2005. *Diccionario panhispánico de dudas (DPD)*. Madrid: Santillana.

Real Academia Española & Asociación de Academias de la Lengua Española. 2010. *Ortografía de la lengua española*. Madrid: Espasa.

Samper Padilla, José Antonio. 1998. "Criterios de edición del léxico disponible: Sugerencias". En: *Lingüística* 10, 311-333.

Sinclair, Stéfan & Rockwell, Geoffrey. 2016. *Voyant Tools*. Web. Disponible en <http://voyant-tools.org/>.

Swiggers, Pierre. 2009. "La historiografía de la lingüística: Apuntes y reflexiones". En: *RAHL: Revista argentina de historiografía lingüística* 1.1, 67-76.

Venturini, Tommaso & Jacomy, Mathieu & Jensen, Pablo. 2021. "What do we see when we look at networks: Visual network analysis, relational ambiguity, and force-directed layouts". En: *Big Data & Society*. January 2021. Disponible en este enlace.

Título / Title

Corpus y herramientas digitales para el estudio de la historiografía lingüística hispánica
Corpus and digital tools for the study of Hispanic linguistic historiography

Resumen / Abstract

Este trabajo se enmarca en el campo de las Humanidades Digitales, entendidas como un área emergente e interdisciplinar en la que convergen las disciplinas humanísticas y las tecnologías digitales. Nuestro objetivo es proponer una forma de acercarnos al estudio de la historiografía lingüística hispánica a través de determinadas herramientas digitales con objeto de extraer nuevos objetos teóricos que puedan ser de reflexión historiográfica y metahistoriográfica. Para ello, mediante herramientas como *Voyant Tools*, *Gephi* u otras más convencionales como *Microsoft Excel*, realizaremos algunos análisis-meta a partir del denominado BiTe_Corpus (Battaner & Herranz-Llácer & Segovia 2021), que contiene los resúmenes de la *Bibliografía temática de historiografía lingüística española: fuentes secundarias* (Esparza *et al.* 2008). Tras la descripción del corpus y la exposición de diferentes análisis, reflexionaremos acerca del potencial para la investigación de estos nuevos objetos de estudio que ofrecen estas herramientas digitales y de sus posibilidades en la investigación historiográfica.

This article is articulated in the field of Digital Humanities, understood as an emerging and interdisciplinary area in which humanistic disciplines and digital technologies converge. The aim is to propose a way of approaching the study of Hispanic linguistic historiography through certain digital tools to extract new theoretical objects that can be of historiographic and metahistoriographic reflection. To this end, using tools such as *Voyant Tools*, *Gephi* or other more conventional tools such as *Microsoft Excel*, it will be carried out some meta-analyses based on the so-called BiTe_Corpus (Battaner & Herranz-Llácer & Segovia 2021), which contains the summaries of the *Bibliografía temática de historiografía lingüística española: fuentes secundarias* (Esparza *et al.* 2008). After the description of the corpus and the presentation of different analyses, we will reflect on the potential for the investigation of these new objects of study offered by these digital tools and their possibilities in historiographical research.

Palabras clave / Keywords

Humanidades Digitales, BiTe_Corpus, Lingüística de corpus, Historiografía lingüística hispánica, Metahistoriografía.

Digital Humanities, BiTe_Corpus, Corpus Linguistics, Hispanic Linguistic Historiography, Metahistoriography.

Código UNESCO / UNESCO Nomenclature

550614, 570100, 570104

Información y dirección del autor / Author and address information

Elena Battaner Moro
Cristina V. Herranz-Llácer
Ana Segovia Gordillo
Departamento de Artes y Humanidades
Facultad de Artes y Humanidades
Universidad Rey Juan Carlos
Campus de Fuenlabrada
Despacho 128, Departamental I
Camino del Molino 5
28943 Fuenlabrada, Madrid
Correo electrónico: elena.battaner@urjc.es; cristina.herranz@urjc.es; ana.segovia@urjc.es