Scriptum Digital
Vol. 6 (2017), pp. 69-82

ETIQUETADO GRAMATICAL Y LEMATIZACIÓN EN EL *CORPUS HISTÓRICO JUDEOESPAÑOL* (*CORHIJE*): PROBLEMAS, SOLUCIONES Y RESOLUCIONES*

AITOR GARCÍA MORENO (*ILC-CSIC* / *IUMP*)
aitor.garcia@cchs.csic.es
F. JAVIER PUEYO MENA (*College of the Holy Cross*)
javier.pueyo@gmail.com

RESUMEN
Tras un breve repaso de las características más sobresalientes del *Corpus Histórico Judeoespañol - CORHIJE* —a modo de recordatorio, pues ya fue presentado en la III edición del *Congreso de Corpus Diacrónicos en lenguas Iberorrománicas* (*CODILI*) en 2014 en Zurich—, mostraremos el proceso de lematización y etiquetado gramatical que se está llevando a cabo sobre el mismo, pasando revista a los distintos problemas detectados y a las soluciones aplicadas durante el mismo que, en algunos casos, nos han obligado a tomar resoluciones, relativamente arbitrarias, en función de los objetivos de descripción y análisis perseguidos: *problemas*, *soluciones* y *resoluciones* que amplifican el título de nuestra presentación.
PALABRAS CLAVE: Corpus lingüísticos, diseño de corpus electrónicos, judeoespañol, diacronía.

LEMMATIZATION AND GRAMMATICAL ANNOTATION OF THE *CORPUS HISTÓRICO JUDEOESPAÑOL* (*CORHIJE*): PROBLEMS, SOLUTIONS, AND RESOLUTIONS

ABSTRACT
After a brief review of the most salient features of the *Corpus Histórico Judeoespañol - CORHIJE* —which was already presented at the 3rd Edition of the *Congreso de Corpus Diacrónicos en lenguas Iberorrománicas* (*CODILI,* Zurich 2014)—, this paper describes the ongoing process of lemmatization and grammatical annotation of the corpus. We focus on describing the challenges we have encountered during the annotation process and the solutions we have applied to them, which, in some cases, have led us to take relatively arbitrary resolutions in accordance with the description and analysis goals we were trying to achieve: *problems, solutions, and resolutions* that amplify the title of our presentation.
KEY WORDS: Linguistic Corpora, Digital Corpus Design, Judeo-Spanish, Diachrony.

Si cualquier corpus diacrónico presenta no pocos retos a la hora de anotar lingüísticamente los diferentes estadios y registros de la lengua a lo largo del tiempo, la variedad judeoespañola añade los suyos propios, debido a su falta de unificación en cada período histórico tanto en el apartado gráfico (textos aljamiados en grafía hebraica o cirílica y textos latinados con diferentes ortografías), como en el léxico o en el morfológico. A esto se añaden su alto grado de diversidad dialectal y la dispersión geográfica de sus hablantes; y es que el judeoespañol presenta en cada período sincrónico una destacable variación gráfica, fonética y morfológica, y se caracteriza —en el ámbito léxico— por la incorporación de préstamos de lenguas diversas, como el hebreo, el turco, el francés, el neogriego o las lenguas eslavas, entre otras.

* Este trabajo se enmarca dentro del proyecto de investigación «Sefarad, siglo XXI (2017-2020): edición y estudio filológico de textos sefardíes» (ref. núm. FFI2016-74864-P), financiado por el Ministerio de Economía, Industria y Competitividad (MINECO).

Fecha de recepción: 25/1/2017
Fecha de admisión: 8/7/2017
ISSN: 2017-640X

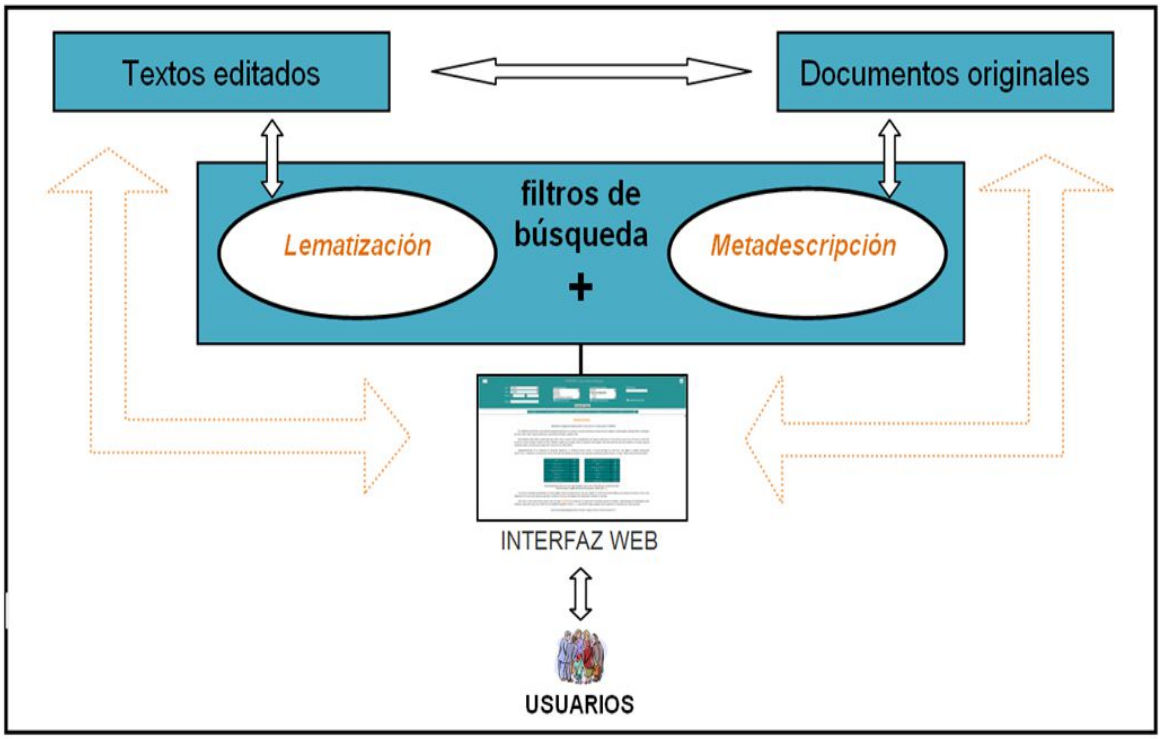
Scriptum Digital
Vol. 6 (2017), pp. 69-82

1. EL *CORHIJE* 1.0

Tal y como se recoge en la página web de nuestro proyecto (García Moreno y Pueyo Mena 2012-2017), el *Corpus Histórico Judeoespañol (CORHIJE)*[1] es un corpus lingüístico accesible en línea, representativo de la evolución de la lengua sefardí, y que está concebido tanto para el investigador como para el lector curioso en general, por su carácter adicional de colección documental. Desde su interfaz web se pueden efectuar búsquedas lingüísticas complejas sobre un número creciente de ediciones críticas de textos sefardíes de todos los lugares, géneros y épocas, así como acceder a la reproducción fotográfica de los documentos originales.

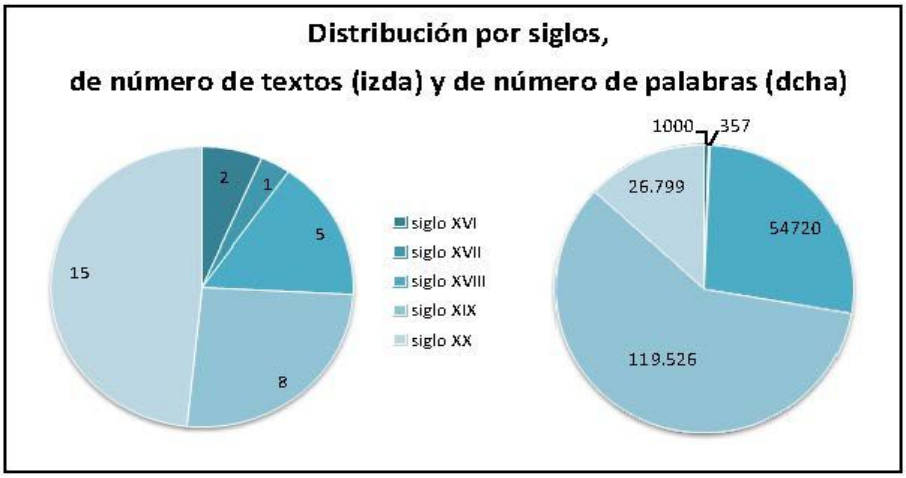
El primer prototipo, obra de F. Javier Pueyo Mena, se desarrolló entre enero de 2012 y diciembre de 2013 en el marco del proyecto *Towards a Representative Corpus of the History of Judeo-Spanish (the* CORHIJE*)* que englobaba a diversos equipos de investigación nacionales y extranjeros, y que fue financiado por la Vicepresidencia de Relaciones Internacionales del CSIC por medio de la ayuda I-LINK-0324.

Como podemos apreciar en el siguiente diagrama, la estructura general del *CORHIJE* combina elementos propios de las colecciones documentales digitales (zona derecha) con otros habituales de los corpus lingüísticos (zona izquierda):

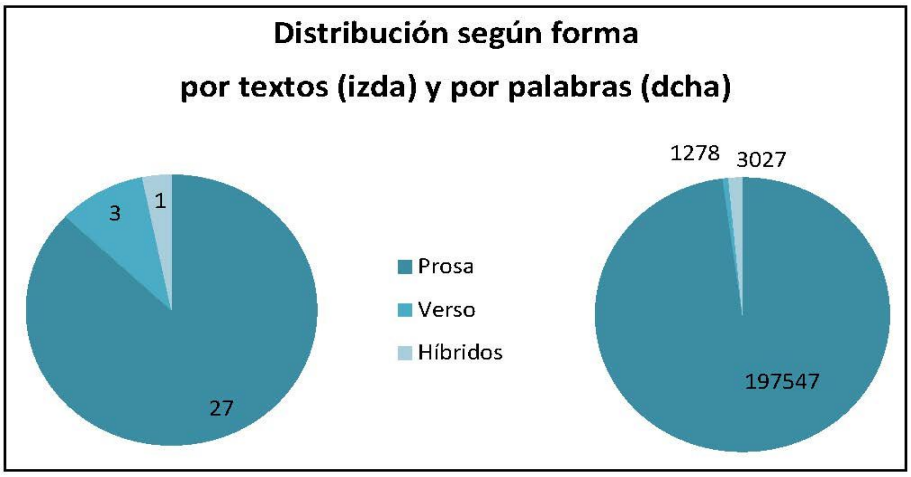


En cifras, el *CORHIJE* contiene actualmente **31 textos** (con un total de **201.852 palabras**) que van desde unas pocas líneas (con 82 palabras el más corto) a las 200 páginas (con 38.430 palabras el más largo) distribuidos por siglos como sigue:

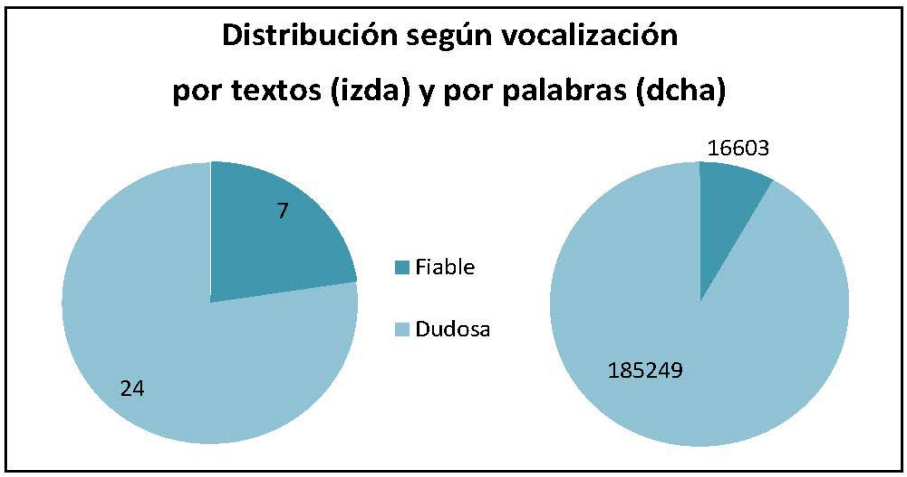
[1] Fecha de última consulta: 15 de enero de 2017.



De estos, **27** son textos en prosa (con 197.547 palabras en total), **3** son textos en verso (1.278 palabras) y **1** (de 3.027 palabras) tiene carácter híbrido por corresponder a un libro de refranes.



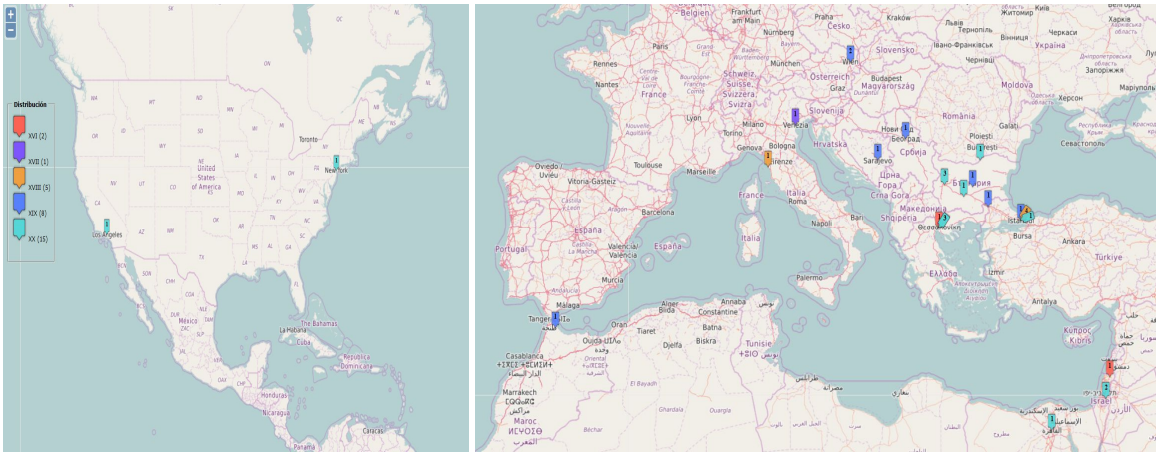
Asimismo, cabe destacar que **7** de ellos ofrecen una vocalización fiable por corresponder a la transcripción de textos orales, textos aljamiados vocalizados, textos escritos en caracteres cirílicos o textos en escritura latina.



Scriptum Digital
Vol. 6 (2017), pp. 69-82

Por último, desde el punto de vista geográfico —para el que se tienen en cuenta tanto el lugar de publicación o producción del texto como el lugar de origen de su responsable (autor, editor, manuscriba, etc.) cuando se conoce— las ciudades representadas actualmente en el corpus, por orden alfabético, son las siguientes: Adrianópolis/Edirne, Belgrado, Bucarest, Constantinopla/Estambul, El Cairo, Esmirna, Filipópolis/Plovdiv, Jerusalén, Kazanlâk, Liorna, Los Ángeles, Nueva York, Ruse/Rustchuk, Safed, Salónica, Sarajevo, Sofía, Tetuán, Venecia, Viena y Xanthi.

El siguiente mapa muestra dicha distribución por lugares, según el número de textos asociados a cada localización y a los siglos de composición:



2. HACIA EL *CORHIJE* 2.0

A partir de transcripciones normalizadas de los textos en aljamía hebraica o cirílica —mediante el uso de diacríticos sobre la grafía hispánica estándar que indican la realización propia del sefardí sobre la base del sistema ideado por Hassán (1978) y, en el caso de los textos latinados, respetando sus propios usos ortográficos pero adaptando la acentuación y puntuación a la ortografía hispánica—, el desarrollo de lo que podríamos denominar *CORHIJE* 2.0 añade nuevas capas de marcación lingüística (lematización y etiquetado gramatical), mapas para la proyección de los resultados de las consultas y, por supuesto, un motor de búsqueda y de obtención de resultados que permite hacer uso de dichas capas de anotación y visualizarlas correctamente.

Para dar cuenta de las peculiaridades del judeoespañol en sus diversas variantes gráficas, y atender a sus variedades diacrónicas y geográficas, decidimos utilizar para el procesamiento lingüístico de los textos la herramienta *Freeling* (Padró 2011), un conjunto de recursos y programas informáticos para el Procesamiento del Lenguaje Natural, que ya se había adaptado con éxito para la anotación lingüística de textos diacrónicos del español transcritos de forma semi-paleográfica en el HSMS (Sánchez Marco, Boleda y Padró 2011).

Freeling está pensado para el análisis multilingüe, utilizando varios módulos encadenados (*tokenizer*, *splitter*, analizador morfológico, etiquetador, desambiguador, etc.) que se sirven de los siguientes elementos para su funcionamiento:

(1) **vocabularios y otros recursos léxicos** (de términos comunes, de nombres propios, de abreviaturas, de locuciones, etc.).

(2) **reglas de dos tipos:**
   (a) de segmentación textual (que le indican a *Freeling* qué secuencias debe considerar como frases, palabras o caracteres alfanuméricos).
   (b) de análisis lingüístico (por ejemplo, para el reconocimiento de partículas enclíticas, de morfemas derivativos o de terminaciones verbales).

(3) **un modelo probabilístico** generado a partir del análisis estadístico de un corpus de entrenamiento, y que se utiliza en los procesos de desambiguación de homónimos y en la asignación de un análisis final para aquellas formas desconocidas que no hayan recibido ninguno a partir de los recursos y reglas descritos anteriormente.

Esta arquitectura permite el uso de diccionarios y de otros recursos semánticos adaptados a cada lengua objetivo, así como el desarrollo de aquellas reglas lingüísticas que reflejen las características particulares de dichas lenguas.

2.1. Etiquetado

El proceso de adaptación de *Freeling* al judeoespañol pasa por crear, modificar, aumentar o eliminar recursos y reglas en los ámbitos señalados. Afortunadamente, hacerlo desde una lengua cercana al judeoespañol como es el español moderno facilita mucho las cosas, ya que, por ejemplo, la variante de *Freeling* para el español está ya diseñada para tener en cuenta una conjugación verbal rica, la presencia de elementos enclíticos, procesos de morfología derivativa, y una sintaxis hispánica en buena parte coincidente con la del judeoespañol: interrogativas, posición de adjetivos, concordancia de género, etc.

A. En el ámbito de los recursos léxicos, ha sido necesaria la creación de un diccionario electrónico a partir de los lemas actualmente incluidos en el *Diccionario Histórico del Judeoespañol - DHJE* (García Moreno 2008-2017)[2] y su ampliación, mediante el desarrollo de un flexionador de formas nominales y de un conjugador verbal, a partir de los infinitivos recogidos en el *DHJE*. El resultado es un diccionario inicial de *formas* / *lemas* / *etiquetas* de alrededor de 300.000 entradas que, en el caso de las formas verbales se han obtenido mediante el uso de un conjugador verbal del español estándar, cuyo análisis final se ha adaptado, a través de reglas sistemáticas, a las terminaciones y cambios de raíz propios del judeoespañol, teniendo en cuenta todas las variantes con diacríticos, como en los siguientes ejemplos:

cant-é > *cant-í*
perd-iste > *perd-ites*

Las etiquetas usadas para describir la información lingüística de cada palabra siguen el estándar EAGLES para lenguas europeas, si bien en determinados casos ha sido necesario modificarlas o crear otras nuevas para describir de forma más precisa algunas formas del judeoespañol. Por ejemplo, para la anotación del participio activo de presente se creó la etiqueta:

cantán CANTAR VM**PP**0S0

[2] 18.585, según la última consulta, realizada el día 15 de enero de 2017.

o para los adverbios en diminutivo:

demaśiyadico DEMASIADO RG**D**

B. En lo relativo a las reglas específicas para los diferentes módulos de procesamiento de *Freeling*, a saber: (a) reglas de división de palabras o *tokenizer*; (b) reglas de división en frases o *splitter*; (c) reglas de afijación, y d) reglas para la detección de expresiones multi-palabra, destaca el hecho de que en nuestro caso debíamos enfrentarnos, por un lado, al uso de guiones medios y bajos, utilizados en la transcripción normalizada para marcar, respectivamente, formas de escritura unitaria en la ortografía del español estándar que en judeoespañol tienden a notarse separadamente como ליב'י מינטי / *leve-mente*, y formas complejas de escritura unitaria en judeoespañol que en español estándar se notan separadamente, como מולו דיירון / *mo_lo dieron*; y, por otro, con no pocos casos de expresiones (principalmente) hebreas insertas en algunos de los textos, y que representan casos claros de *cambio de código*, y que por lo tanto debían ser ignoradas por *Freeling*.

C. En cuanto a las reglas lingüísticas que se han creado o modificado, podemos citar la incorporación de clíticos propios del judeoespañol (con y sin diacríticos) y su reconocimiento en posiciones sintácticas ajenas a las del español estándar. Por ejemplo, se desarrollaron reglas para dar cuenta de clíticos que no existen en español (*lis*, *si*, *sin*, *mos/moś*) o para anotar casos no existentes en español moderno, como los clíticos pospuestos a una forma verbal conjugada (*díjolis*)[3]:

1	2	3	4	5	6	7	8	9	10
Clitic to erase from word form	**Affix to add to the resulting root**	**Expected category**	**Tag for suffixed**	**Check lemma adding accents**	**Enclitic suffix (special accent behaviour in Spanish)**	**Prevent Freeling from assigning more tags to the word**	**Lemma/Form/Root/Affix to assign**	**Consider the suffix always, not only for unknown words**	**How to retokenize the word if necessary**
lis	*	^V	*	1	1	0	L	1	$$+les:$$+PP
si	*	^V	*	1	1	0	L	1	$$+se:$$+PP
sin	*	^V	*	1	1	0	L	1	$$+se:$$+PP
mos	*	^V	*	1	1	0	L	1	$$+nos:$$+PP
moś	*	^V	*	1	1	0	L	1	$$+nos:$$+PP

Por último, se tuvieron en cuenta todas las posibles combinaciones y posiciones del sistema de clíticos sefardí, por lo que formas complejas como *púśomoslo* (esp. *nos lo puso*) o *dijéronselo* (esp. *se lo dijeron*) se analizan correctamente a partir de una sola

[3] De acuerdo con la [Columna 1] le decimos a *Freeling* que en las secuencias terminadas en *-lis*, extraiga la cadena que le antecede (*díjo-*). [Columna 2] El asterisco le indica a *Freeling* que no debe quitar ni añadir nada a la forma *dijo* para continuar con el análisis (esto puede ser necesario en otros casos). [Columna 3] Si el resultado es cualquier verbo (conjugado o no: ^V) entonces *lis* puede y debe analizarse como clítico. [Columna 4] No se aplica a los clíticos. [Columnas 5 y 6] Los 1 y 0 son indicadores para *Freeling* sobre si debe intentar encontrar *dijo* tanto con tildes como sin ellas (en este caso es importante ya que la cadena con tilde *díjo* no existe en el diccionario, pero sí existe la forma sin tilde). [Columna 7] No se aplica a los clíticos. [Columna 8] La L indica que se debe asignar el mismo lema al conjunto *díjolis* que a *dijo*, esto es, DECIR. [Columna 9] Esta regla debe aplicarse tanto a palabras conocidas como a desconocidas. [Columna 10] Qué análisis (además de *Verbo*) debemos asignar al conjunto. En este caso, el mismo que a *les* (es decir: *Pronombre Personal, tercera persona del plural*).

entrada léxica y de una regla lingüística única que los describe a todos. El análisis final de *púśomoslo* resultaría en la siguiente anotación:

Forma	Lemas	Etiquetas
púśomoslo	poner+nos+lo	VMIS3S0+PP1CP000+PP3CNA00

La ampliación de reglas lingüísticas ha sido necesaria también, por ejemplo, para algunos casos particulares de afijación (diminutivos, aumentativos, etc.) del español sefardí. Así, la versión del español estándar de *Freeling* no contempla, por ejemplo, derivados en *-ico* (también escrito *-iko*), muy productivos en judeoespañol, por lo que hubo que añadir una regla específica para su reconocimiento. *Freeling* tampoco contempla que a partir de un adverbio puedan derivarse diminutivos, algo relativamente frecuente en judeoespañol, por lo que una vez creada la regla para *-ico*/*-iko*, hubo que ampliarla para que pudiera funcionar también con dichos adverbios, como vemos en la siguiente tabla para el tratamiento de la forma *demaśiyadico*[4].

1	2	3	4	5	6	7	8	9	10
Affix to erase from word form	**Affix to add to the resulting root**	**Expected category**	**Tag for suffixed**	**Check lemma adding accents**	**Enclitic suffix (special accent behaviour in Spanish)**	**Prevent Freeling from assigning more tags to the word**	**Lemma/Form/Root/Affix to assign**	**Consider the suffix always, not only for unknown words**	**How to retokenize the word if necessary**
ico	o	^RG	RGD	0	0	0	L	0	-

2.2. Lematización

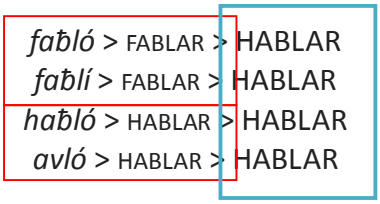
En lo que corresponde a la lematización de *CORHIJE*, de una parte, se decidió hacerla compatible con la de otros corpus diacrónicos hispánicos, por lo que el lema de cada forma se introduce en el diccionario principal sin diacríticos (aunque sí con tildes ortográficas) y se elige la forma que mejor puede reflejar la correspondiente al español actual, cuando esta existe.

Sin embargo, para no dejar sin reflejar los fenómenos de variación (principalmente fonéticos) de carácter sistemático en español sefardí —los cambios en el timbre vocálico (incluido el cierre de vocales medias en posición átona), los casos de aféresis, prótesis, metátesis, epéntesis y paragoge, las asimilaciones y disimilaciones, etc.— y permitir así proyectarlos sobre los mapas dialectales que se producen al realizar consultas sobre el corpus, se decidió crear un nivel superior de lematización que los recogiera, por lo que cada *forma* recibe un LEMA y una especie de ARCHILEMA.

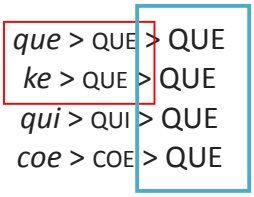
Mientras los LEMAS nos permiten agrupar las formas flexivas de una palabra y salvar las diferencias derivadas de usos ortográficos particulares, los ARCHILEMAS nos permiten agrupar las distintas variantes del contínuum dialectal sefardí, como entre:

[4] Al quitar *-ico* [Columna 1], nos queda la raíz *demaśiyad-* por lo que, de acuerdo con la [Columna 2], la regla añade una *-o* antes de buscar en el diccionario *demaśiyad*. [Columna 3]: si el diccionario indica que *demaśiyad+o* es un adverbio (^RG) entonces, [Columna 4] el conjunto es analizado como adverbio diminutivo (RGD) y se le asigna [Columna 8] el mismo lema que a *demaśiyado*; en este caso, DEMASIADO.

Scriptum Digital
Vol. 6 (2017), pp. 69-82



en el caso de palabras con flexión, o



en el caso de palabras invariables.

Para dichos LEMAS y ARCHILEMAS se ha optado por una representación gráfica sin diacríticos que sigue la del español estándar cuando esta existe. Cuando no existe en español, como en el caso de ciertos préstamos, se ha estandarizado la forma judeoespañola a una de ortografía hispánica por dos motivos:

(1) para abstraer un lema que represente a todas las formas tanto diacrónicas como dialectales del judeoespañol, y
(2) para permitir la integración de dicha lematización con otros corpus hispánicos y facilitar así el uso del *CORHIJE* a los hispanistas no sefardistas.

Pero lo cierto es que, al menos en el nivel LEMA, el uso de diacríticos podría ampliar el número de variantes cubiertas por los distintos ARCHILEMAS.

2.3. Desambiguación

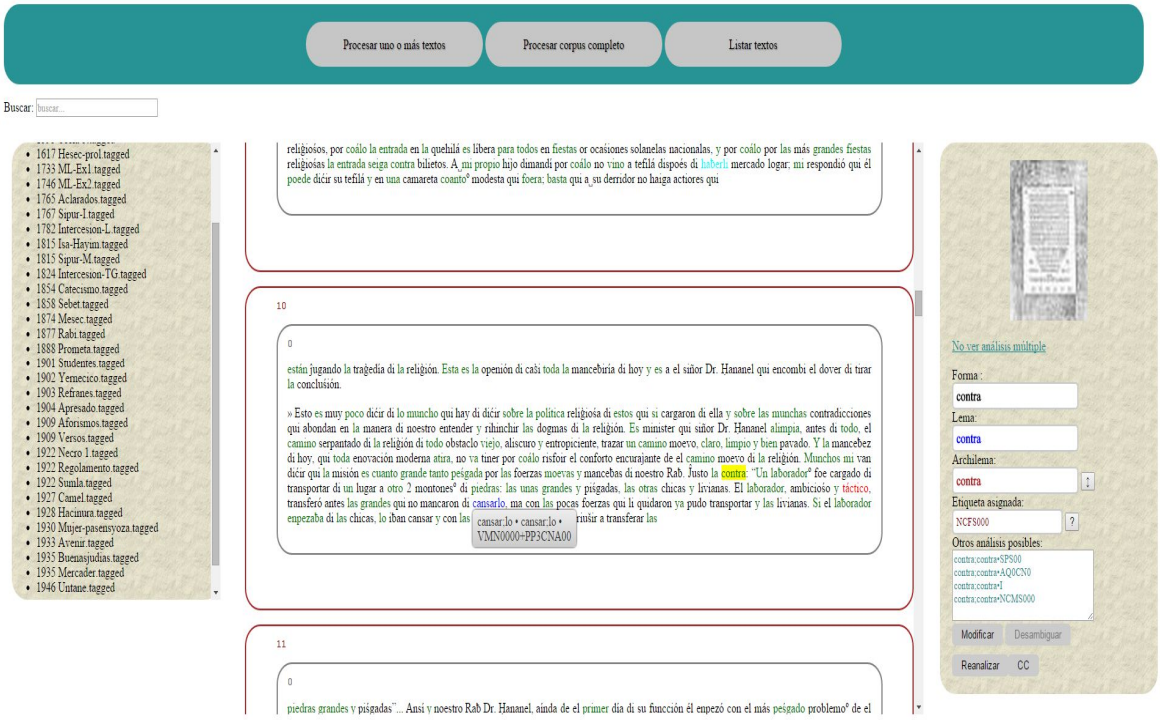
El modelo probabilístico de *Freeling* que mencionábamos anteriormente está lógicamente ligado a un análisis estadístico previo de las características lingüísticas y de su frecuencia en un corpus determinado. Otra de las ventajas de nuestra adaptación de los recursos de *Freeling* para el español y de su uso sobre un corpus del judeoespañol, es precisamente el aprovechamiento de la ingente información estadística que posee *Freeling* para el español y que utiliza a la hora de desambiguar formas homónimas o de etiquetar las desconocidas, según su contexto sintáctico y sus patrones morfológicos.

Si bien es cierto que entre el español y el judeoespañol hay diferencias tanto sintácticas —por ejemplo, algunas estructuras propias que podríamos asimilar a ciertas «pseudocopulativas» como *Mošé merece matado* (= 'Mošé merece morir')— como morfológicas —por ejemplo, formas verbales no personales construidas sobre el tema de perfecto como *quisendo* (*vs.* esp. *queriendo*) o *tuvido* (*vs.* esp. *tenido*), o adverbios en *-mente* construidos sobre formas no adjetivales, como *cercamente* (esp. *cercanamente*) o *verdadmente* (esp. *verdaderamente*)—, hay que tener en cuenta que en nuestro plan de actuación, el uso estadístico de *Freeling* se circunscribe a una primera fase de marcación que está siendo manualmente corregida.

De acuerdo con nuestra experiencia previa, entendimos que partir de un corpus previamente analizado y con una alta precisión se consideró preferible a marcar manualmente desde el comienzo cientos de miles de formas desconocidas.

Scriptum Digital
Vol. 6 (2017), pp. 69-82

Una vez creados todos los recursos y realizada la adaptación de *Freeling* al judeoespañol, esta se integró en una herramienta propia de análisis textual denominada **CORHIJE-APP** que, partiendo del texto plano de las transcripciones, devuelve un texto en formato XML con toda la información lingüística incorporada, pero que mantiene todas las características textuales de la obra para facilitar su lectura y su posterior presentación en los resultados de las consultas[5]. La herramienta permite de forma muy sencilla incorporar formas desconocidas por *Freeling* (que se señalan en rojo), y revisar tanto los análisis generados automáticamente por las reglas lingüísticas creadas (en azul) como los análisis asignados a formas homónimas potencialmente ambiguas (en verde).

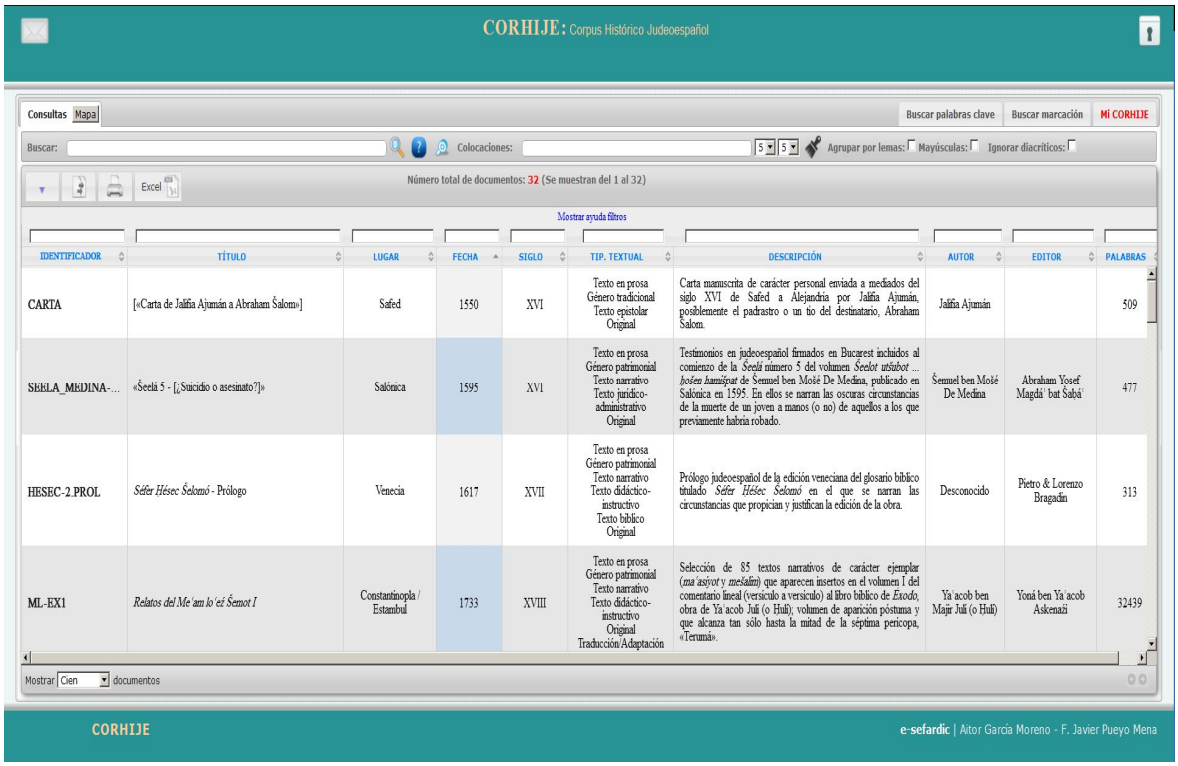


Tal y como hemos señalado anteriormente, la corrección semi-manual de una parte representativa del corpus (en cuanto a distribución geográfica, cronológica, gráfica y de tipología textual) que actualmente estamos llevando a cabo, permitirá establecer un corpus fiable de entrenamiento, con el cual se generarán nuevos modelos probabilísticos que reflejen con mayor exactitud los textos sefardíes y que permitirán a *Freeling* realizar el proceso de desambiguación automática de formas homónimas —con múltiples etiquetas y lemas posibles— y proponer categorías gramaticales para las formas desconocidas.

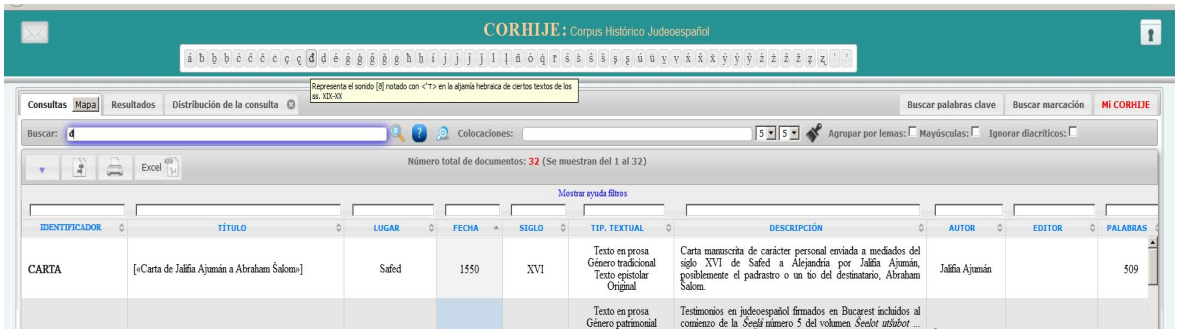
[5] Desde un punto de vista técnico, usamos: 1) XML y HTML para la representación de los textos y sus marcas, 2) *Javascript* y *Perl* para la programación y ejecución de Freeling en el *background* y 3) *Sqlite* como base de datos.

3. *CORHIJE* 2.0 - EL PROTOTIPO

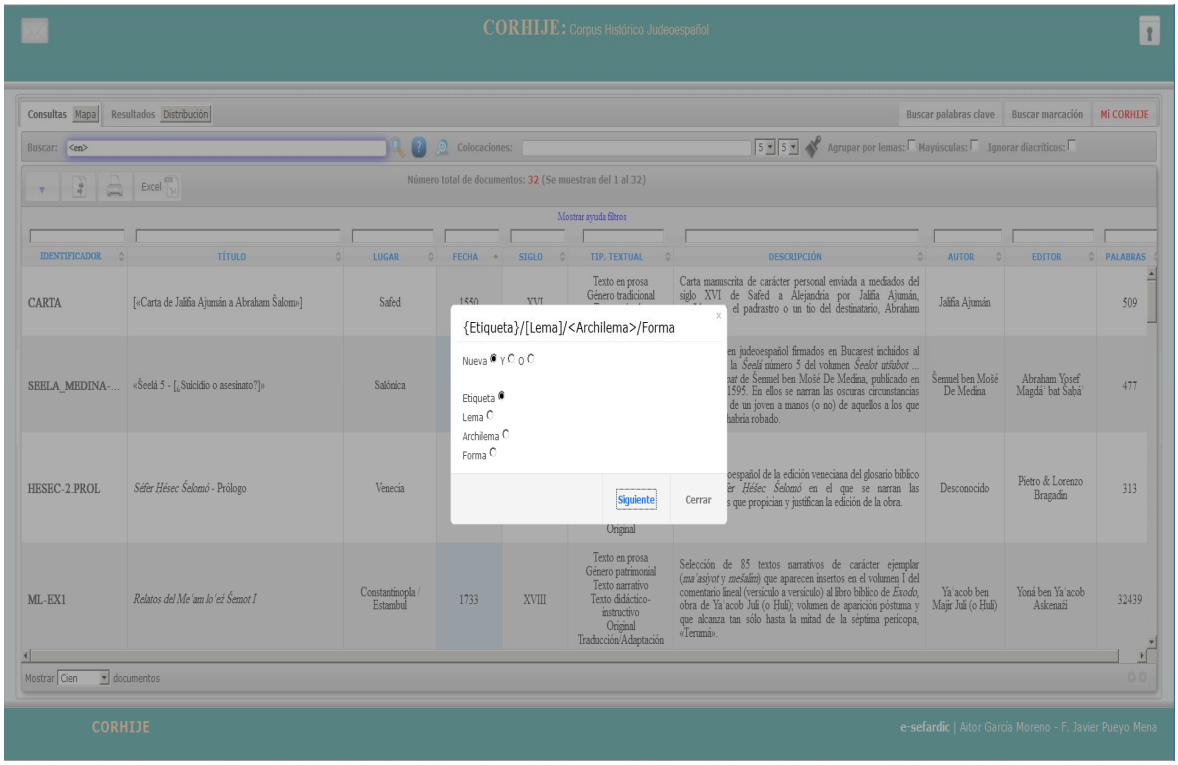
Tras la lematización, el etiquetado y la desambiguación iniciales del corpus, se ha creado un prototipo de buscador y visualizador de resultados en la web.



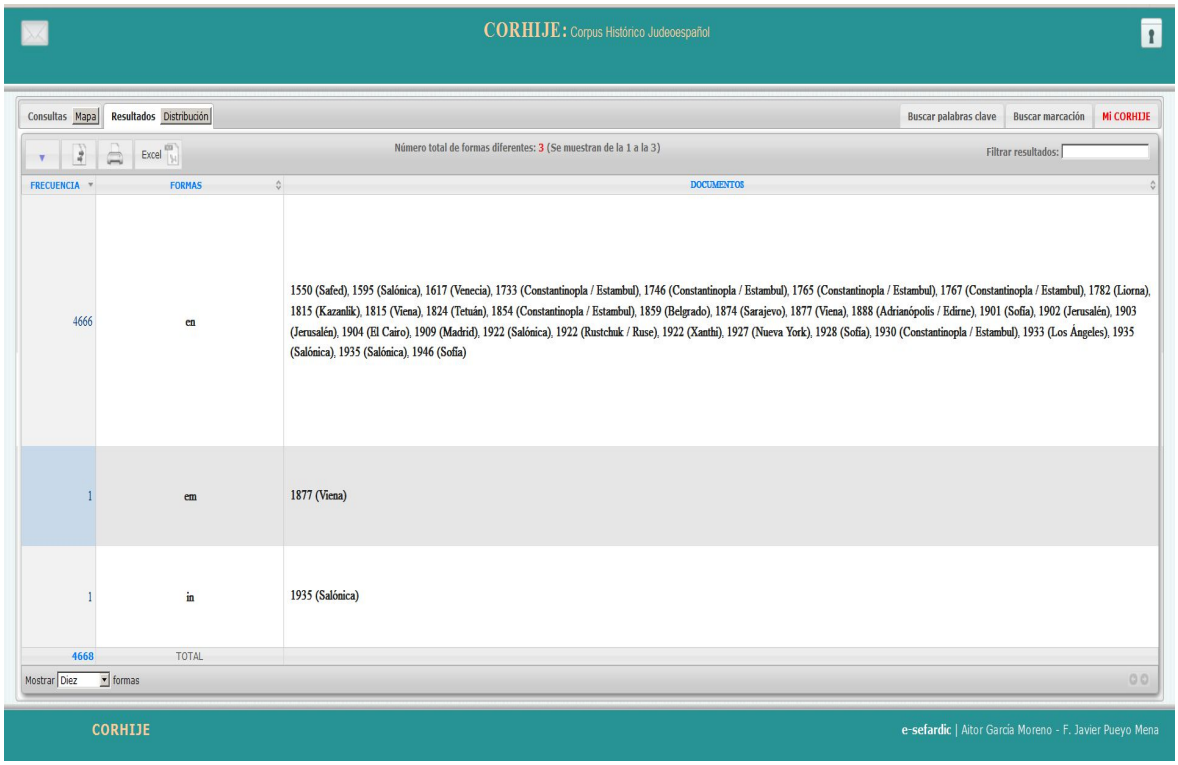
Este nuevo buscador permite llevar a cabo consultas, simples o combinadas, por forma, lema, archilema y etiqueta gramatical, todas ellas con la posibilidad de introducir comodines. Para la inserción de signos diacríticos —que también pueden ser obviados si se prefiere— el acceso a la caja de búsqueda despliega un teclado virtual con ayudas contextuales en cada caso.



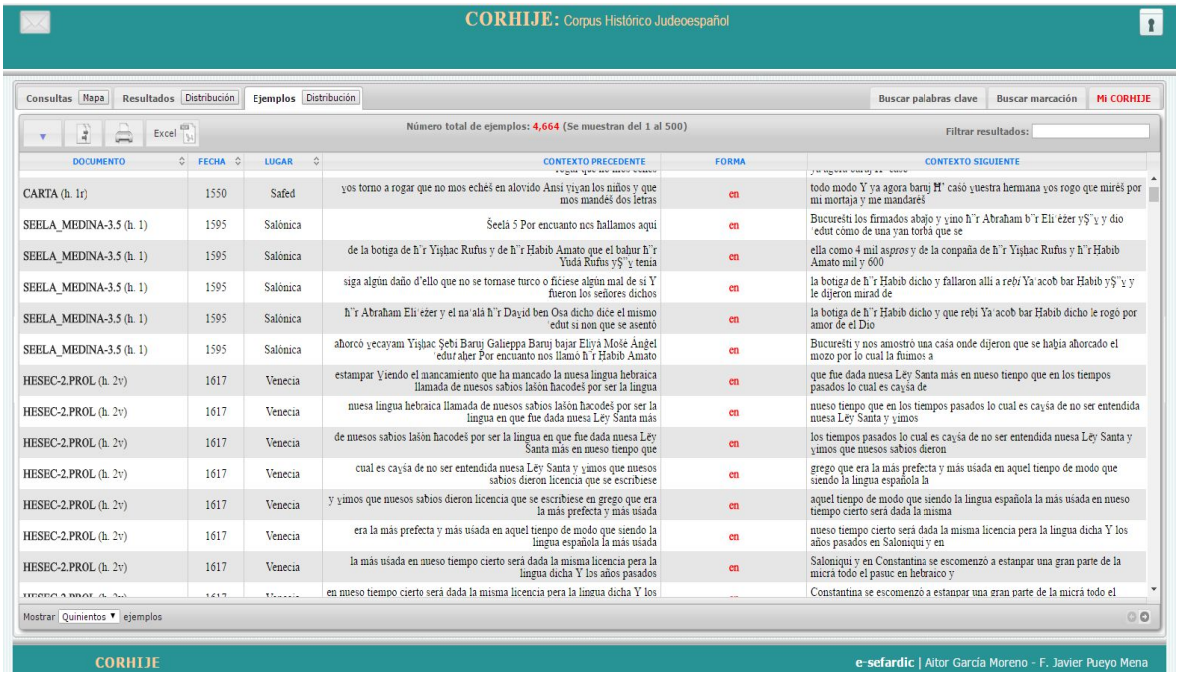
Para facilitar la elaboración de las consultas más complejas —sobre todo las referidas a etiquetas gramaticales—, se ha creado un «asistente» por pasos.



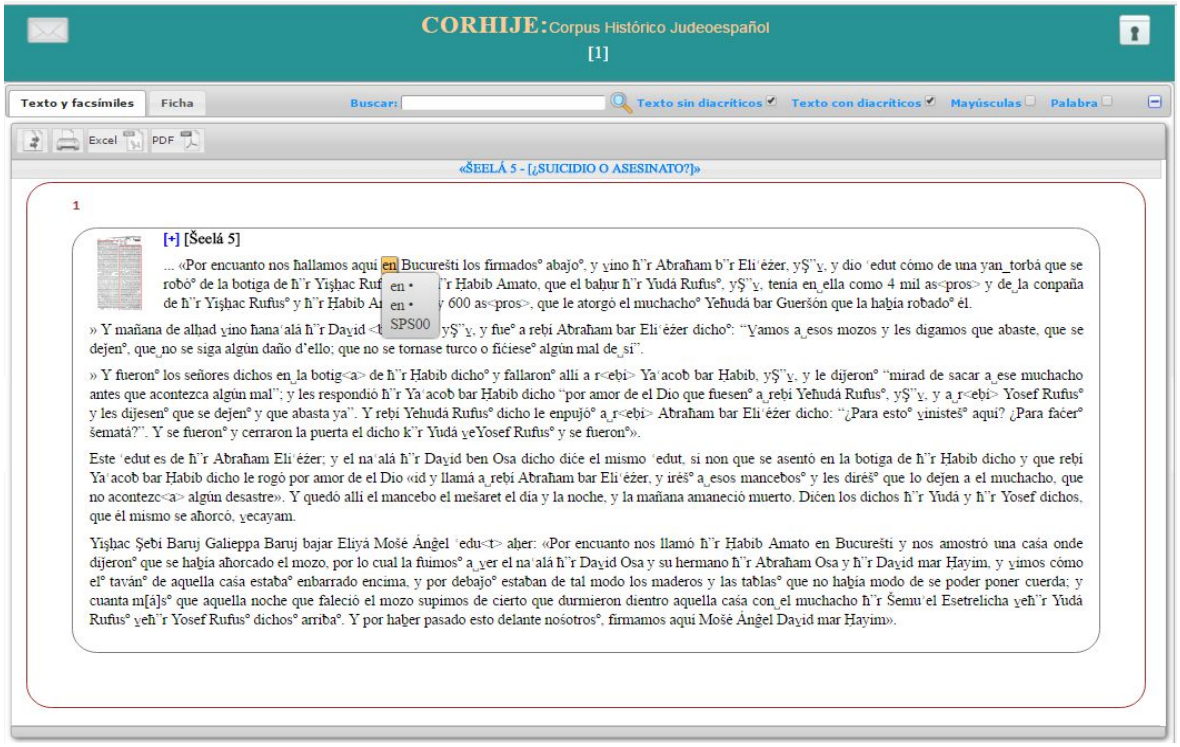
Los resultados se presentan en forma de un listado de formas que cumplen las condiciones de la consulta realizada, tal y como vemos en la siguiente búsqueda correspondiente al archilema <EN>:



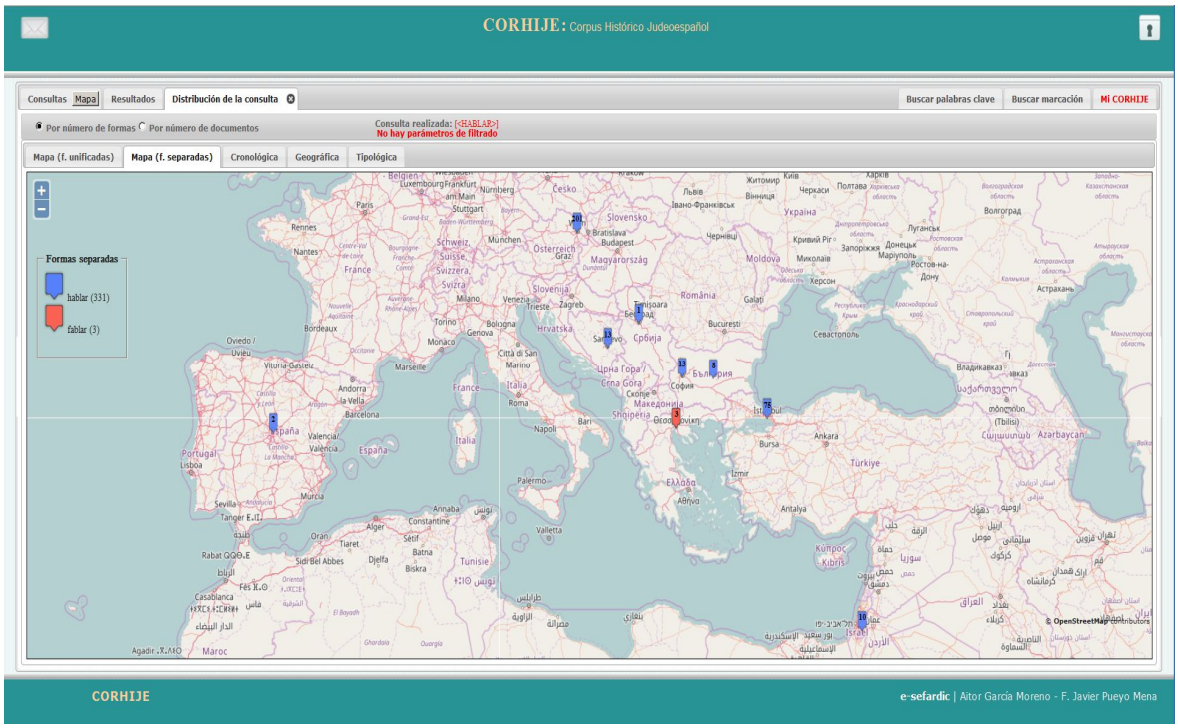
Una vez seleccionada una de las formas propuestas (*en*/*em*/*in*), se presenta un listado contextual de la misma (KWIC) con todos los ejemplos encontrados en el corpus, como se muestra en la siguiente figura:



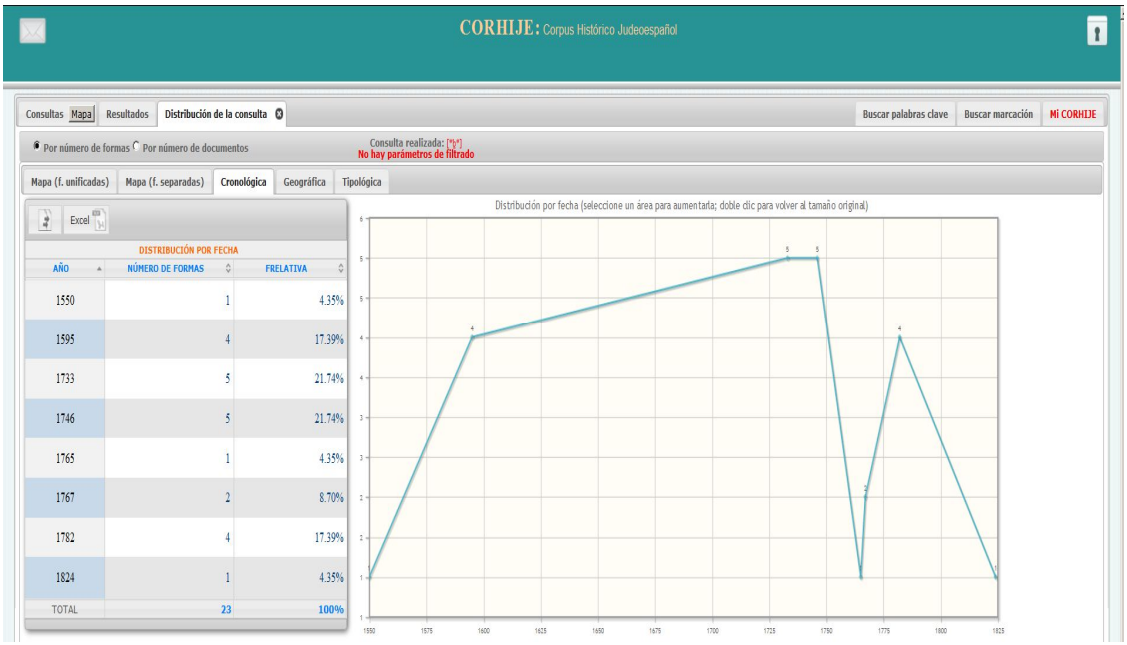
siendo posible también consultar el texto completo correspondiente a cada ejemplo para su mayor contextualización o para cotejarlo inmediatamente con el facsímil:



Asimismo, los resultados pueden ser proyectados en mapas, como en el siguiente ejemplo de las distintas formas documentadas del archilema <HABLAR>, con la opción de agrupación *por lemas* (*hablar/fablar*) en lugar de la opción por defecto que es la agrupación *por formas* (*avló/haƀlí/faƀlaremos*, etc.):



Al mismo tiempo se generan gráficas estadísticas de distintos tipos, como la que puede observarse en el siguiente ejemplo, que corresponde al progresivo abandono del uso de <ו> *vav* para /v/ en formas romances actualmente notadas con <b> en judeoespañol:



Funcionalidades —algunas de ellas todavía en desarrollo—, que, junto a la implementación completa del visor de documentos con la versatilidad ya probada en *CORHIJE* 1.0, habrán de ocupar nuestro tiempo, al menos, hasta la próxima edición de *CODILI*.

Entre tanto, confiamos en que nuestro sistema de *doble lematización* inspire a todos aquellos corpus lingüísticos –sea cual sea su orientación– donde la variación desempeña un papel especialmente destacado; pues entendemos que el establecimiento de distintos niveles de agrupación (ya sea formal, como en nuestro caso, o de otro tipo) permite dar un mejor y más variado tratamiento a los datos lingüísticos que cada corpus maneja.

REFERENCIAS BIBLIOGRÁFICAS

GARCÍA MORENO, Aitor (dir.) (2008-2017): *Diccionario Histórico del Judeoespañol (DHJE)*. http://www.esefardic.es/dhje [Consulta: 30/6/2017]

GARCÍA MORENO, Aitor y Francisco Javier PUEYO MENA (2013-2017): *Corpus Histórico Judeoespañol (CORHIJE)*. http://www.esefardic.es/corhije [Consulta: 30/6/2017]

HASSÁN, Iacob (1978): «Transcripción normalizada de textos judeoespañoles», *Estudios Sefardíes*, 1, pp. 147-150.

PADRÓ, Lluís (2011): «Analizadores Multilingües en FreeLing», *Linguamática*, 3, 2, pp. 13-20. http://www.linguamatica.com/index.php/linguamatica/article/view/115 [Consulta: 1/8/2017]

SÁNCHEZ MARCO, Cristina, Gemma BOLEDA and Lluís PADRÓ (2011): «Extending the Tool, or How to Annotate Historical Language Varieties», in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Stroudsburg (PA): Association for Computational Linguistics, pp. 1-9. http://dl.acm.org/citation.cfm?id=2107637&CFID=979433322&CFTOKEN=43121501 [Consulta: 1/8/2017]