Lematización automática y diccionarios electrónicos*

(Automatic lemmatisation and electronic dictionaries)

Lleal Galceran, Coloma
Univ. de Barcelona. Fac. de Filología. Dpto. de Filología Hispánica. Sección de Lengua Española. Gran Vía de les Corts Catalanes, 585
08007 Barcelona

BIBLID [1137-4454 (2006), 21; 331-343]
Recep.: 08.10.04
Acep.: 18.10.05

Descripción del proceso de elaboración de un diccionario con la ayuda de los medios que facilita la informática: caracterización del corpus y de los criterios de selección y de transcripción; análisis de la estructura de la base de datos y de los fundamentos de los programas de lematización automática; descripción del proceso lexicográfico y avance de los resultados.

Palabras Clave: Lexicografía. Lematización. Diccionario electrónico. Diacronía. Siglo XV.

Hiztegi bat informatikak bideraturiko baliabideen laguntzaz osatzeko prozesuaren deskripzioa: corpusaren eta hautapen irizpideen karakterizazioa, datu basearen egitura eta lematizazio automatikorako programak aztertzea, prozesu lexikografikoaren deskripzioa eta emaitzen aurrerapena.

Giltza-Hitzak: Lexikografia. Lematizazioa. Hiztegi elektronikoa. Diakronia. XV. mendea.

Description du processus d'élaboration d'un dictionnaire avec l'aide des moyens fournis par l'informatique: caractérisation du corpus et des critères de sélection et de transcription; analyse de la structure de la base de données et des bases des programmes de lématisation automatique; description du processus lexicographique et devancement des résultats.

Mots Clés: Lexicographie. Lematisation. Dictionnaire électronique. Diachronie. XV^{ème} siècle.

* El presente estudio ha sido posible gracias a la ayuda del Ministerio de Ciencia y Tecnología, Proyectos de investigación números PB1998-1223 y BFF2002-00898.

1. PRESENTACIÓN

Desde hace varios años, el *Grupo de Historia y Contacto de Lenguas* (GHCL) está trabajando en la confección de un diccionario del castellano del siglo XV. Partimos de textos escritos o publicados en la Corona de Aragón porque, desde un primer estudio que realizamos sobre el lenguaje cancilleresco[1], constatamos la importancia de estos textos en la configuración de la lengua renacentista.

Recordemos que hacia mediados del siglo XV la monarquía aragonesa conquistó Nápoles y el monarca con su corte se estableció en esa ciudad. Los contactos entre Nápoles y las grandes ciudades de la Corona (Valencia, Barcelona, Zaragoza…) fueron constantes y, sobre todo, los intercambios entre los intelectuales de la época. De ahí que los usos lingüísticos renacentistas penetrasen profundamente en la lengua culta de la época, tanto en el catalán como en el castellano –ambas lenguas oficiales de la Corona, tras el proceso castellanizador del aragonés iniciado en esa época[2]– y actuasen como modelo de un nuevo estilo, cuya influencia sobre el castellano inmediatamente posterior no puede ser negligida. Por ello, el estudio de los textos escritos en castellano en esta zona peninsular merecía una especial atención.

Partíamos también de la hipótesis de que el nuevo estilo lingüístico penetró profundamente en todos los niveles de la lengua[3], por lo que no podíamos limitarnos al estudio de los textos literarios. Como consecuencia de todo ello, iniciamos el proceso de recopilación de materiales lingüísticos a partir de los cuales podríamos elaborar nuestro diccionario.

Nuestro corpus actual consta de dos tipos fundamentales de textos: textos no literarios (A) y textos literarios (B), cada uno de ellos subdividido en: A-1: textos administrativos, cancillerescos y jurídicos; A-2: textos científicos; B-1: textos narrativos e históricos y B-2: textos novelescos; con una extensión de cerca de un millón y medio de formas, distribuidas de forma regular en cada una de las subdivisiones.

Dada la extensión de nuestro corpus, vimos la necesidad de incorporar las innovaciones metodológicas que presenta el tratamiento automático de los textos. Porque, en efecto, en una época en que el estudio del vocabulario medieval y renacentista con la ayuda de los medios informáticos es ya una realidad, no nos podemos mantener al margen de esta aportación. Pero, al mismo tiempo, pretendíamos que nuestro trabajo no se viera sometido a

---

1. LLEAL, Coloma (1997). *Vocabulario de la Cancillería aragonesa (siglo XV)* y *El castellano del siglo XV en la Corona de Aragón*. Zaragoza: Institución "Fernando el Católico".

2. LLEAL, Coloma (2001). «Historia de la lengua e historia de la lengua literaria a la luz del catalán de los siglos XVI y XVII», *Epos*, XVII, 89-106.

3. LLEAL, Coloma (1995). «El secretario, el nuncio y la difusión del latinismo en el siglo XV», *Lletres Asturianes,* 56, 19-34.

las imposiciones que tan a menudo parten del mundo de la informática y que nos hiciera olvidar nuestra labor como filólogos. Por ello emprendimos el diseño de un sistema sencillo, a partir de un programario existente en el mercado y de fácil manejo, que adaptamos convenientemente a nuestras necesidades[4]. Voy a intentar, en las líneas que siguen, resumir las principales etapas de nuestro trabajo[5].

2. PROCESO DE SELECCIÓN Y TRANSCRIPCIÓN DE LOS TEXTOS

Partimos siempre de textos originales, algunos de ellos manuscritos y otros en ediciones de la época. En todos los casos, hemos procurado ser rigurosamente fieles al original introduciendo el menor número posible de modificaciones: separación con un punto volado de las formas aglutinadas y unión de los componentes de los nombres propios, entendidos como una única unidad lingüística. En cualquier caso, siempre es posible recuperar la forma original. Asimismo, se incluye entre paréntesis cuadrados la referencia del texto y del folio correspondiente al inicio de cada sección, a fin de permitir la posterior ubicación automática de cada una de las formas. Pero hemos evitado el uso de una etiquetación exhaustiva previa, que poco añade al conocimiento del texto, y en cambio complica considerablemente el tratamiento inicial.

> [B1-CroAra-123-r] tanta gracia se fauoreçia / y se acompañaua que todo lo real se vencia de su valer. amaua en·demasia la reyna las cosas que el rey amaua. Y conociendo que allende el amor del rey su virtud y mereçimientos lo adebdauan y requerian / suplico al rey su señor / que diesse conclusion al matrimonio que se tractaua del excellente infante su hermano con la illustre doña Guillerma_de_Muncada: fija del noble / magnifico / egregio don Gaston_de_Muncada: vizconde de Bearn que tenia solo en Catalueña trezientos caualleros. Concluydo el matrimonio por la diligencia que puso en·ello la reyna: quedo el infante y su esposa mas todo el linage tan principal de Muncada mucho aficionado al seruicio de·la reyna.

Fig. 1. Tratamiento inicial del texto

4. El diseño del programa ha sido facilitado al equipo de la Universidad de Salamanca que dirige la profesora Mª Nieves Sánchez, que lo utiliza satisfactoriamente en sus estudios lexicográficos.

5. LLEAL, Coloma (2002). «Una base de datos para el estudio del léxico del siglo XV», en Echenique, Mª T. y Sánchez, J. (eds.), *Actas del V Congreso Internacional de Historia de la Lengua Española*, II, Madrid, Gredos, 2201-2210.

ANGLADA ARBOIX, Emilia (2003). «Un diccionario general y etimológico del castellano del siglo XV en la Corona de Aragón», comunicación presentada en el *VI Congreso Internacional de Historia de la Lengua Española* (Madrid, octubre de 2003). Se publicará en las *Actas* correspondientes, en prensa.

ANGLADA ARBOIX, Emilia (2004). «Un banco de datos electrónico: a propósito de la confección de un diccionario del castellano del siglo XV en la Corona de Aragón», comunicación presentada en el *I Congreso Internacional de Lexicografía Hispánica*. Se publicará en las *Actas* correspondientes, en prensa.

Este texto mínimamente modificado es exportado a una base de datos, con lo que podremos obtener ordenaciones de los elementos de que consta a partir de distintos criterios[6]. Esta base de datos consta, inicialmente, de dos campos: uno para las formas y otro para la situación. Ello nos va a permitir iniciar la tarea lexicográfica.

Situación	Forma
B1-CroAra-142v	el
B1-CroAra-142v	rey
B1-CroAra-142v	de
B1-CroAra-142v	Aragon
B1-CroAra-142v	que
B1-CroAra-142v	le
B1-CroAra-142v	mandaria
B1-CroAra-142v	dar
B1-CroAra-142v	por
B1-CroAra-142v	esposa
B1-CroAra-142v	a
B1-CroAra-142v	doña
B1-CroAra-142v	Violante de Hurrea
B1-CroAra-142v	fija
B1-CroAra-142v	del
B1-CroAra-142v	noble
B1-CroAra-142v	don
B1-CroAra-142v	Johan Ximenez de Hurrea
B1-CroAra-142v	donzella
B1-CroAra-142v	noble
B1-CroAra-142v	y
B1-CroAra-142v	d·
B1-CroAra-142v	especial
B1-CroAra-142v	fermosura

Fig. 2. La base inicial

3. PROCESO DE LEMATIZACIÓN

3.1. No nos interesaba trabajar exclusivamente con las formas, sino con los lemas a partir de los cuales confeccionar nuestro diccionario. Además, dada la multiplicidad de variantes que encontramos en todo texto medieval o renacentista, queríamos poder contar con un listado de variantes para facilitar las búsquedas posteriores. Así, las distintas formas flexivas de un verbo se agrupan bajo un mismo lema –que podríamos considerar "canónico" según la

6. Las ventajas de trabajar con bases de datos, por la ductilidad del sistema, son de sobras conocidas. Precisamente por ello, quisiera hacer constar aquí nuestra admiración y respeto por quienes nos han precedido y, con medios materiales muy rudimentarios, nos han legado obras de inestimable valor.

norma culta de la época–, pero se especifican las variantes correspondientes –piénsese, por ejemplo, en el lema *hazer*, que puede presentar las variantes *hazer, hacer, fazer, faser, far*–. Por ello, introducimos tres nuevos campos en nuestra base: uno para los lemas o vocablos, otro para las variantes y otro para la función. Y, a continuación, confeccionamos un programa que automáticamente procede a la lematización del texto.

3.2. La primera cuestión que debíamos plantearnos era, precisamente, la de las funciones que asignábamos a las formas. No voy a reproducir aquí la discusión acerca de los problemas relacionados con la categorización, porque no se trataba tanto de formular teorías como de establecer una norma que nos permitiese aplicar una "etiqueta" a las formas del texto. De acuerdo con ello, partimos inicialmente de las siguientes clases de palabras: sustantivos, adjetivos, verbos, adverbios, pronombres personales, demostrativos, posesivos, indefinidos, relativos, identificadores, locativos y numerales, preposiciones, conjunciones, interjecciones y nombres propios[7]. Estas etiquetas iniciales serán revisadas y completadas en una etapa posterior, con el texto ya lematizado. Así, en el caso de las formas pronominales, distinguiremos, por ejemplo, entre pronombres demostrativos con valor sustantivo o con valor adjetivo. En el programa inicial, esta distinción solo podrá hacerse para ciertas formas como *esto, eso, aquello*, siempre con valor sustantivo, mientras que las demás se marcarán, inicialmente, con la etiqueta "pron. dem. adj.", etiqueta que posteriormente, en la etapa de análisis del texto, deberá ser revisada en función del contexto. Y de manera similar actuamos con el resto de las formas pronominales. Asimismo, en el caso de los verbos se marcará posteriormente el carácter auxiliar, transitivo, intransitivo o pronominal. Algunas de estas subclasificaciones podrán hacerse automáticamente (en el caso de la mayoría de verbos auxiliares, por ejemplo), pero otras deberán establecerse manualmente en una etapa posterior de revisión del texto.

3.3. Partimos, en primer lugar, de la constatación de que en la lengua hay un número relativamente reducido de formas gramaticales, pertenecientes a inventarios cerrados, que presentan, en cambio, un alto índice de lectura (adverbios, preposiciones, conjunciones, artículos, pronombres…). Piénsese que en nuestro caso solo 138 formas gramaticales constituyen algo más del 50% del total del texto, relación que varía muy poco de un texto a otro[8].

---

7. Partimos, básicamente, de la clasificación de Alcina-Blecua, 1989 y que, en líneas generales, coincide con la que aplican los grupos de investigación CLiC y TALP, del Departamento de Lingüística de la Universitat de Barcelona y del Departamento de Lenguajes y Sistemas Informáticos de la Universitat Politècnica de Catalunya (véase CIVIT TORRUELLA, Montserrat. *Criterios de etiquetación y desambiguación morfosintáctica de corpus del español*, Barcelona: SEPLN, 2003).

8. MULLER, Charles (1968): *Estadística lingüística*. Madrid, Gredos, 1973.

LÓPEZ MORALES, Humberto (1983): «Lingüística estadística», en López Morales, H. (ed.), *Introducción a la lingüística actual*, Madrid, Playor, 209-225.

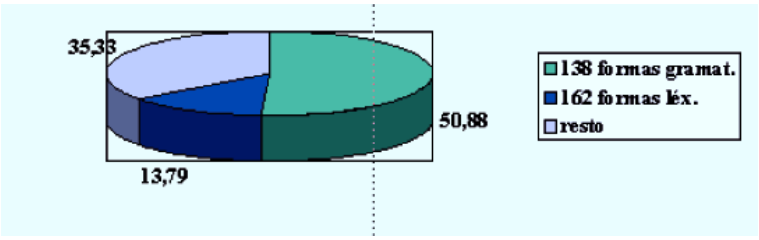

Fig. 3. Índice de lectura de las 300 formas más frecuentes

Este carácter constante del índice de lectura de las formas gramaticales, así como el hecho de pertenecer a inventarios cerrados, nos permite establecer un catálogo completo a partir del cual podamos programar las instrucciones adecuadas para proceder a su lematización. Se trata, en suma, de describir la gramática básica del funcionamiento de estas formas. Asimismo, podemos establecer un catálogo de las formas léxicas fundamentales, formado por un reducido número de unidades léxicas de carácter muy repetitivo, dependientes en este caso del tipo de texto, pero que pueden ser fácilmente identificadas mediante un listado de frecuencia de las formas. Con ello tendremos lo que podríamos denominar el *lexicón* mínimo[9] a partir del cual podremos lematizar más del 60% del texto, lo cual constituye, de entrada, un porcentaje nada desdeñable.

Forma	Fr_abs
de	40422
y	32005
la	27379
el	21001
e	18874
que	18055
los	17623
en	15710
por	15011
de	14405
que	12558
et	12487
del	10425
se	8914
a	8863
las	8803
a'	7595
con	7584
en'	7580
no	7458
o	6264
mas	6233
le	5616
su	5206

Fig. 4. Listado de frecuencia de las formas

9. PICOCHE, Jacqueline (1991): «Lexicographie théorique et appliquée: un dictionnaire des mots de haute fréquence», *Cadernos de Lingua*, 3, 87-110.

Las instrucciones de lematización para las unidades gramaticales son de dos tipos, según el carácter invariable o flexivo de las formas que nos ocupan. En el primer caso, bastará con indicar que en el campo correspondiente al lema o *vocablo* y en el campo de las *variantes* hay que copiar el mismo contenido que se halla en el campo *forma*. Del tipo: "copia la *forma* en los campos *vocablo* y *variante* y escribe "prep." en el campo *función* siempre que en el campo *forma* aparezca "*a*" o *"de"* o *"por"* o *" con"...* etc." En otras ocasiones, habrá que prever la existencia de variantes distintas de un único lema: piénsese en el adverbio "*assi*", que puede aparecer con las formas "*assy*", "*asy*" "*asi", "ansi"* o *"assi*". Finalmente, en las formas flexivas (artículos, pronombres...) habrá que dar cuenta de las formas a partir de las cuales se manifiesta el lema correspondiente, con una instrucción del tipo: "escribe "*el, la lo*" en los campos *vocablo* y *variante* y "*art*." en el campo *función* siempre que en el campo *forma* se encuentre "*el*" o "*la*" o "*lo*" o "*las*" o "*los*". Obsérvese que en esta etapa inicial lematizamos siempre de acuerdo con la función de mayor frecuencia. En etapas posteriores se podrán precisar, de acuerdo con el contexto, muchas de estas categorizaciones.

3.4. En cuanto a las formas léxicas, de inventario abierto, hemos visto que, por una parte, podíamos incluir en el *lexicón* previo aquellas formas que presentan un alto índice de frecuencia. Pero para el resto, deberemos aplicar una serie de instrucciones que parten de su estructura morfológica.

Así, a partir de los morfemas derivativos y flexivos podemos lematizar un número considerable de estas formas léxicas. Para identificarlas debidamente, partimos de un listado inverso de las formas que nos permite constatar las regularidades y también las excepciones: así, todas las formas terminadas en {-ble} serán, en principio, adjetivos, o las terminadas en {-cion} o en {-ura} serán sustantivos femeninos, o las que presenten el segmento final en {-miento} serán sustantivos masculinos. Bastará, en este caso, introducir una instrucción del tipo: "siempre que encuentres una *forma* terminada en 'ura', copia la *forma* en los campos *vocablo* y *variante* y escribe "sust. fem." en el campo *función*." Cuando constatamos la existencia de posibles excepciones, la instrucción correspondiente habrá de dar cuenta de ellas: así, en el caso de las formas en {-ble} observamos la presencia del verbo *hable/fable* o del sustantivo *condestable*; entre los adjetivos en {-oso} encontramos los sustantivos femeninos *cosa* y *rosa*; entre las formas terminadas en {-ura} hallamos el adjetivo femenino *dura*; entre las formas en {-mente} no son adverbios el sustantivo femenino *mente*, ni el adjetivo *clemente* o el verbo *tormente*.

Para los paradigmas verbales, aparte del listado de verbos de alta frecuencia (*ser, estar, haver, dezir, hazer, ir*...), a menudo con numerosas irregularidades, que habremos incluido en el *lexicón* inicial, disponemos de un listado de desinencias verbales regulares a partir de las cuales podemos deducir el lema correspondiente. Así, siempre que encontremos una forma terminada en {-assemos} podemos formular una instrucción del tipo: "cuenta el número de letras de la *forma*, elimina las seis últimas letras y escribe el resto+"r" en el campo *vocablo"*. El número de formas que puede lematizarse

a partir de reglas de este tipo es considerable. Y, también aquí, el listado inverso de las formas nos ayuda a detectar las excepciones (pienso, por ejemplo, en la terminación {-amos}, generalmente morfema de primera persona plural, pero que también puede aparecer como secuencia final de los sustantivos *balsamos*, *clamos* o *ramos*).

En el análisis de las terminaciones verbales, el orden de aplicación de las instrucciones es fundamental para evitar la generación de vocablos inexistentes: si aplicamos primero la regla correspondiente a {-amos} generaremos vocablos como **pagariar, *contentariar, *abreviariar*. Por tanto, primero habrá que pensar en identificar las formas del condicional {-ariamos} y solo después podremos pasar a las del presente, es decir, habrá que lematizar en primer lugar las formas cuyo morfema verbal contenga un mayor número de caracteres.

Con todo, difícilmente podremos diferenciar, solo por la forma de la desinencia, las formas de presente de indicativo de los verbos de la primera conjugación de las formas de presente de subjuntivo de los verbos de la segunda y tercera (*amamos* vs *temamos*), con lo que probablemente generaremos vocablos del tipo **temar* que requerirán una posterior corrección.

3.5. Un tercer grupo de reglas parte de la estructura sintáctica de las frases. Por una parte, constatamos la existencia de numerosas coincidencias formales entre determinadas formas verbales y las formas de sustantivos y adjetivos, que solo podrán diferenciarse por el contexto. Veamos un ejemplo: formas aparentemente idénticas en su segmento final como *canta, manta,* o *santa*, presentan posibilidades combinatorias diferenciadas. Así, mientras es posible *la canta, lo canta, el canta* (dada la ausencia de tilde en los textos de la época), *las canta, los canta*, solo podremos encontrar *la manta* o *la santa*, pero no **el manta, *lo manta, *las manta, *los manta* o bien **el santa, *lo santa, *las santa, *los santa*. La posibilidad de alternancia del elemento precedente, lematizado inicialmente como artículo por presentar mayor frecuencia de uso en esta función, nos permite categorizar la forma *canta* como verbo y corregir la función de las formas *el, la lo, los, las* en esta posición como pronombres personales. Asimismo, *canta* presenta la posibilidad de combinarse con *le*, cosa que no ocurre en el caso de *manta* o *santa*. Además, dado que en el tratamiento inicial del texto hemos desaglutinado las formas verbales con pronombre enclítico, separándolas mediante un punto volado, en el texto podemos encontrar también la forma *canta·*, pero no **manta·* o **santa·* Por otra parte, el sustantivo solo presenta variación de número, mientras que el adjetivo presentará variación de género y número. La aplicación combinada de estas normas nos permite identificar un número considerable de formas verbales y nominales, así como las formas de artículo de las pronominales.

A partir de criterios similares podemos diferenciar los frecuentes casos de homonimia: *fuera* verbo *ser* (seguido de adjetivo) o verbo *ir* (seguido de la preposición *a*), de *fuera* adverbio; *que* conjunción, de *que* relativo; *si* pronombre personal, de *si* conjunción condicional… también podemos identificar las formas verbales compuestas, así como las locuciones y perífrasis, que consignaremos en el campo *variantes*.

3.6. La última operación del programa lematizador consiste en la corrección automática de los errores más frecuentes. Piénsese, por ejemplo, en las alternancias entre el diptongo [jé] o [wé] en posición tónica y la vocal simple [e] y [o] en posición átona en numerosas formas verbales que puede haber generado lemas del tipo *pueder* o *tiener*, que pueden corregirse automáticamente. O bien en la existencia de numerosas alternancias gráficas en los textos de la época que puede generar lemas distintos (*haver, aver, auer, haber*; *hazer, fazer, faser, azer, hacer; hablar, fablar, faular…*) que, sin embargo, deberían aparecer unificados. En este caso, hemos incluido en nuestro programa una instrucción *ad hoc* que mantiene las alternancias en el campo *variante*, pero las unifica en el campo *vocablo*. Y, finalmente, en la existencia de variantes puramente gráficas (grafías dobles de <f> o <p> en alternancia con consonantes simples; grupo <nm> en alternancia con <mm>, <mpm>, <npm>; alternancia entre <yn>, <in> o <jn>, y similares), que también deberán ser unificadas en el campo *vocablo* y conservadas en el campo *variante*.

3.7. Terminada esta operación, con el texto lematizado casi en su totalidad (actualmente nuestro programa lematiza automáticamente un 96,7% de las formas), deberemos proceder a una cuidadosa revisión: la total ausencia de intuición y de competencia lingüística de los ordenadores es de todos conocida.

Forma	Vocablo	Función	Variante	Situación
largo	largo -a	adj.	largo -a	B1-Viaje-015v
destierro	destierro	sust. masc.	destierro	B1-Viaje-015v
fueron	ser	verbo perf.	ser (+part.)	B1-Viaje-015v
perdonados	perdonar	verbo trans.	perconar	B1-Viaje-015v
y	y	conj.	y	B1-Viaje-015v
les	el ella ello	pron. pers.	le	B1-Viaje-015v
entrego	entregar	verbo trans.	entregar	B1-Viaje-015v
sus	suyo -a	pron. pos. ad.	su	B1-Viaje-015v
bienes	bien	sust. masc.	bien	B1-Viaje-015v
perdidos	perder	verbo trans.	percer	B1-Viaje-015v
.				B1-Viaje-015v
el	el la lo	art.	el la lo	B1-Viaje-015v
qual	qual	pron. relat. sust.	qual	B1-Viaje-015v
viendo	ver	verbo trans.	ver	B1-Viaje-015v
que	que	conj.	que	B1-Viaje-015v
ya	ya	adv.	ya	B1-Viaje-015v
la	el la lo	art.	el la lo	B1-Viaje-015v
vejez	vejez	sust. fem.	vejez	B1-Viaje-015v
le	el ella ello	pron. pers.	le	B1-Viaje-015v
aquexaua	aquexar	verbo trans.	aquexar	B1-Viaje-015v
por	por	prep.	por	B1-Viaje-015v
bien	bien	adv.	bien	B1-Viaje-015v
conseruar	conservar	verbo trans.	conseruar	B1-Viaje-015v
la	el la lo	art.	el la lo	B1-Viaje-015v
cosa	cosa	loc. sust. fem.	cosa publica	B1-Viaje-015v
publica	publico -a	adj.	publico -a	B1-Viaje-015v
y	y	conj.	y	B1-Viaje-015v
el	el la lo	art.	el la lo	B1-Viaje-015v

Figura 5. El texto lematizado

4. PROCESO DE CONFECCIÓN DEL DICCIONARIO

A partir de la primera base, en la que cada una de las ocurrencias aparece acompañada del lema correspondiente, la función, la variante y la situación, entramos de lleno en la etapa claramente lexicográfica. Para ello, creamos otras bases relacionadas que nos permiten, por una parte, definir el sentido que en cada contexto presentan estas formas y, por otra, introducir información acerca de la etimología del término.

4.1. En primer lugar, para facilitar la labor de comprensión del sentido de los términos, creamos una nueva base en la que añadimos a nuestra base inicial dos nuevos campos: uno que presenta el contexto inmediato de cada una de las formas, con los doce términos precedentes y los doce siguientes, y otro en el que especificamos la acepción que le corresponde a cada ocurrencia. Asimismo, mediante la totalización de los vocablos de esta base, creamos otra, llamada VALOR, que consta de los campos *vocablo* y *función* procedentes de la base anterior (inicialmente con una ocurrencia para cada vocablo), más dos nuevos campos para la *acepción* y el *sentido*. A partir de este momento, trabajamos conjuntamente con las dos bases, CONTEXTO y VALOR, y vamos especificando el sentido de cada una de las ocurrencias de un vocablo, al tiempo que añadimos más registros a medida que encontramos nuevas acepciones de un vocablo:

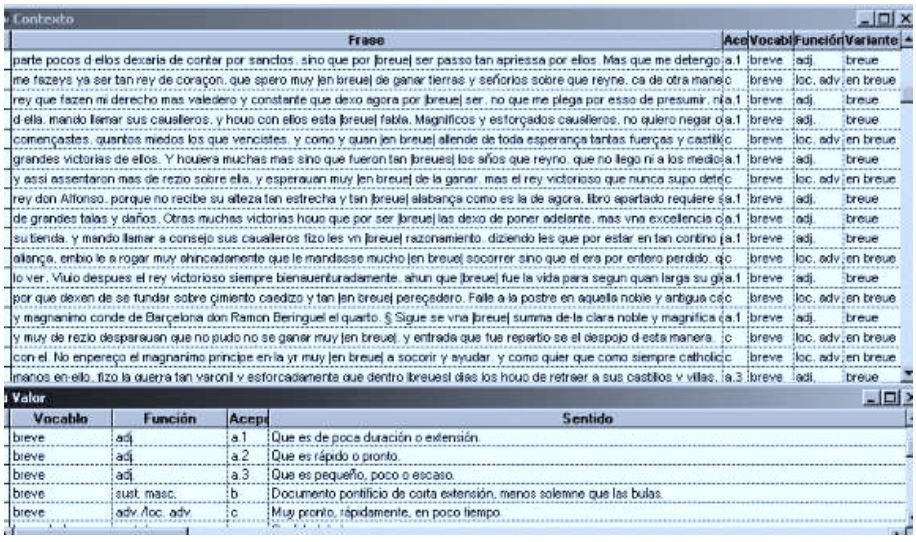

Fig. 6. Especificación de las acepciones en las dos bases relacionadas

4.2. También a partir de la primera base, creamos otra denominada ÉTIMO, en la que tenemos de entrada un campo para los vocablos, y añadimos cinco nuevos campos para el étimo, la base léxica, la fecha de la primera documen-

tación en el DCECH, la fecha de nuestra primera documentación y otro campo para comentarios adicionales. A partir de esta base podremos establecer la etimología de cada uno de los lemas, su cronología, así como el conjunto de términos que parten de una misma base léxica.

5. REUNIÓN DE LOS DATOS

Como resultado final, pretendemos ofrecer un diccionario complejo que facilite datos de naturaleza muy diversa procedentes de las tres bases –CONTEXTO, VALOR y ÉTIMO– con que hemos trabajado.

Para ello, elaboramos una nueva base, que denominamos REUNIÓN, que, tras la aplicación de un programa diseñado para ello, agrupará todos los datos de que disponemos en las bases iniciales. Así, podremos obtener información acerca de:

a) cada uno de los lemas, con su correspondiente etimología y cronología y, en su caso, la especificación de su carácter neológico;

b) acerca de los lemas de nuestro corpus relacionados con una misma base léxica;

c) acerca de todas las variantes y las formas con que se presenta un lema, con especificación del número de ocurrencias de cada uno de ellos;

d) acerca de la frecuencia absoluta y relativa de cada lema, así como de cada una de las acepciones;

e) acerca del tipo de texto en que aparece cada una de las acepciones, con un ejemplo para cada uno de ellos;

f) acerca de los sinónimos de cada acepción en nuestro corpus…

Toda esta información puede ofrecerse en formato convencional y nos proporcionará datos de innegable interés para el estudio diacrónico de la lengua.

fuego Etimología: latín FOCUM 'hogar, hoguera'. Documentación: DCECH: Orígenes. Nebrija: *huego.*
Frecuencia absoluta: 219; frec. relativa: 0,02241 %
Família léxica en el corpus: *fogage, fogoso -a, hogaça, hoguera.*

a. *sust. masc.* Materia en combustión que emite luz y calor.
Frecuencia absoluta: 182 (=86,30%)
Variantes atestiguadas*: fuego* (102), *huego* (80); formas atestiguadas*: fuego* (99), *fuegos* (3), *huego* (80). Distribución: A2: 115 (=63,18%); B1: 56 (=30,77%), B2: 11 (=6,05%).

Ejemplos: *anchos como tres dedos cumplidos o quatro siquiera y tengan cerca muy viuo |fuego| para que los fierros de mucho rusientes bueluan como blancos*. [A2-Albeyt-024-v]; *por ende traspassa todo el calor sin que se vea nada del |huego|. Tardamos en esta misma ciudad despues tres dias por vnas questiones que*… [B1-Viaje-057v]; *las otras damas y donzellas que con ella stauan de·las llamas del |fuego| a·fuerça la quitaron. y luego la reyna con otros caualleros llegaron*… [B2-Grisel-026v].

b. *sust. masc.* Hoguera o incendio.
Frecuencia absoluta: 15 (=6,84%)
Variantes atestiguadas: *fuego* (9), *huego* (6); formas atestiguadas: *fuego* (3), *fuegos* (6), *huego* (3), *huegos* (3). Distribución: B1: 15 (=100%).
Sinónimos: *hoguera.*
Ejemplos: *presentes muy poca parte segun aquellas que se perdieron por sus grandes guerras |fuegos| y derruecos de vnos a otros. nunca se pudo tanto screuir* [B1-Viaje-005v]; *dias y .vij. noches queriendo saber con tal esperiencia que grandes fueron los |huegos| de Troya quando fue presa. mando que matassen la mayor parte de*·[B1-Viaje-136r].

c. *sust. masc*. Quemadura hecha en un tejido orgánico con un hierro candente con fines curativos.
Frecuencia absoluta: 11 (=5,02%)
Variantes atestiguadas: *fuego* (9), *huego* (2); formas atestiguadas: *fuego* (1), *fuegos* (8), *huegos* (2). Distribución: A2: 11 (=100%).
Ejemplos: *Si por ventura el agrion fuere de mucho tiempo den le vnos |fuegos| del traues y luengo con vn subtil fierro mucho quemante.* [A2-Albeyt-043r]; *Empero acaten con diligencia que al cortar ni dando los |fuegos| no se acuesten a·la tetilla de medio la boca por el mucho*… [A2-Albeyt-050v].

d. *sust. masc.* Enfermedad de las caballerías en que se producen erupciones en la piel.
Frecuencia absoluta: 5 (=2,28%).
Variantes atestiguadas: *fuego* (3), *huego* (2); formas atestiguadas: *fuego* (3), *huegos* (2). Distribución: A2: 5 (100%).
Ejemplos: *que se embeua en·la hinchazon. § Cura para quitar el dolor del |fuego| donde lo touiere el cauallo. § Vinagre alquena o alheña vn poco azeyte* [A2-Albeyt-031r]; *agua tan honda que pueda cubrir el dicho daño del todo sobre los |huegos|. Quando saliere echen le ceniza fecha de salzedo siquiere salç.* [A2-Albeyt-043r].

e. *sust. masc.* Ardor o exaltación producida por los sentimientos.
Frecuencia absoluta: 6 (=2,74%).
Variantes atestiguadas: *fuego* (5), *huego* (1); formas atestiguadas: *fuego* (5), *huego* (1). Distribución: B1: 4 (=66,66%), B2: 2 (=33,33%).
Ejemplos: *y que dexasse tan a peligro los reynos que no atajasse el |fuego| tan poderoso de·la començada vnion contra el rey.* [B1-CroAra-082v]; *Dios lo permite por sus pecados abhominables que siempre las fuerças y |fuego| les cresce. Assi que se pueden estos herejes dezir açotes* [B1-Viaje-117r].

Fig. 7. Estructura de una entrada del diccionario

Pero si se ofrece además en formato electrónico, las posibilidades de búsqueda se multiplican considerablemente. Porque, en efecto, será posible obtener también

a) listados de los lemas o de las formas de mayor frecuencia;

b) podrán hacerse búsquedas a partir de las variantes o de las formas, y no solo de los lemas;

c) podrán obtenerse listados de todos los neologismos del siglo XV;

d) o de los lemas agrupados según sus funciones gramaticales, tanto en el conjunto del corpus como en un tipo de texto o en un texto concreto;

e) listados de los lemas en cuya definición aparece determinado rasgo seleccionador –pienso, por ejemplo, en rasgos del tipo "aplicado a las caballerías", que selecciona un grupo de adjetivos–:

Vocablo	Sentido
abierto -a	[Caballería] que separa excesivamente las extremidades al andar.
aguatomado -a	[Caballería] que padece una enfermedad en que se producen abcesos en el interior del casco.
alazan	[Caballería] que tiene el pelo de color rojizo o semejante al de la canela.
anquiseco -a	[Caballería] que tiene las ancas secas y descarnadas.
bayo -a	[Caballería] que tiene el pelo de color blanco amarillento.
bocamuelle	[Caballería] que es excesivamente sensible a los toques del freno.
braig	[Caballería] que tiene el pelo de color mezclado de ruano y rojizo.
diestro -a	[Caballería] que es tirada por la brida por una persona que anda a su lado.
enclavado -a	[Caballería] que hendas en el casco por haber introducido demasiado un clavo al herrarla.
enrastrado -a	[Caballería] que se mueve con dificultad a causa de un enfriamiento.
escalivado -a	[Caballería] que padece una enfermedad intestinal que le provoca desecamiento.
estrellero -a	[Caballería] que anda levantando mucho la cabeza.
peceño -a	[Caballería] que tiene el pelo de color negruzco como la pez.
rossillo -a	[Caballería] que tiene el pelo de color mezclado de blanco, negro y castaño.
ruan -ana	[Caballería] que tiene el pelo de color mezclado de blanco, gris y rojo.
rucio -a	[Caballería] que tiene el pelo de color pardo claro o blanquecino.
sabino -a	[Caballería] que tiene el pelo de color rocado o rojizo.
soberbio -a	[Caballería] que tiene un temperamento fogoso o violento.
topino -a	[Caballería] que tiene cortas las cuartillas y pisa con la parte anterior del casco.
tordillo -a	[Caballería] que tiene el pelo de color mezclado de negro y blanco.
trescanado -a	[Caballería] que tiene una mancha de color claro en la frente.

Fig. 8. Selección de los adjetivos

f) listados de las locuciones o de las perífrasis (identificadas en el campo *variantes*);

g) listados de las colocaciones o de los contextos más frecuentes de un término;

h) listados de todos los contextos en que aparece determinado fenómeno lingüístico…

Y tantos otros que se podrían sugerir, posibles precisamente por el carácter intertextual de un diccionario electrónico. Se trata, en suma, de un sinfín de posibilidades que, sin duda, compensarán el esfuerzo realizado.