El proyecto MORFOGEN. Diseño de una aplicación web para visualizar familias léxicas

María José Rodríguez-Espiñeira
Universidade de Santiago de Compostela
mjose.rodriguez.espineira@usc.es

Yolanda López Arca
Universidade de Santiago de Compostela
yolanda.lopez.arca@usc.es

1. INTRODUCCIÓN

El proyecto que presentamos con el acrónimo MORFOGEN[1] pretende contribuir al estudio de la morfología histórica del español. La hipótesis que lo sustenta es que la clasificación y definición de las palabras derivadas no debe ofrecerse de forma aislada, sino en forma de redes o constelaciones. El estudio de dichas redes puede abordarse desde dos puntos de vista: el diacrónico y el sincrónico, y no tiene por qué haber coincidencia entre las interpretaciones, los análisis y los resultados obtenidos desde ambas perspectivas. El proyecto tiene utilidad desde dos puntos de vista: teórico y aplicado.

1.1. Como se ha señalado, las relaciones derivativas establecidas en un estudio descriptivo sincrónico no coinciden necesariamente con las relaciones derivativas consideradas históricamente. Hacer visibles estas relaciones genéticas es uno de los objetivos de este proyecto. En Pena (2012) puede verse un detallado análisis de la secuencia derivativa Verbo → Nombre deverbal y de los reajustes que se han producido en esta secuencia como resultado de las lagunas existentes en las series latinas y de las nuevas creaciones romances. Para ilustrarlo con un pequeño ejemplo, extraído del mismo texto, los sustantivos *canto, traslado* y *compra* son derivados verbales ya que indican la acción de *cantar, trasladar* y *comprar*, respectivamente (cf. *escuchamos el canto de los ruiseño-*

[1] La participación en el congreso HDH2013 ha sido posible gracias a una subvención de la Xunta de Galicia (*Desarrollo de información en una aplicación web para el estudio morfogenético del léxico*. Ref. 10PXIB204249PR. Dirección Xeral de I + D. Período: 10/08/2010 a 30/09/2013). El proyecto global ha contado también con financiación estatal (*La red de morfología genética en el NDHE*. Ref. FFI2008-03532. Ministerio de Ciencia e Innovación. Período 01/01/2009 al 31/12/2011; *Estudio morfogenético del léxico español*. Ref. FFI2012-38550. Ministerio de Economía y Competitividad. Período: 01/02/2013 a 31/01/2016).

res, el traslado de la corte se produjo en 1808, *la compra de esta casa será un negocio*). Sin embargo, al estudiar en paralelo las series derivativas latinas y españolas, se comprueba que históricamente dicha relación V → N es válida únicamente para *comprar* (documentada en 1095) y *compra* (documentada en 1102[2]). En cambio, el verbo *trasladar* está creado sobre el sustantivo *traslado* (heredado del latín *translātus*, que a su vez es deverbal de *transferre* y este de *ferre*). Por otra parte, el sustantivo *canto* se ha asociado en español con el verbo *cantar* (que continúa el latín *cantāre*, frecuentativo de *canere*), pese a que etimológicamente *canto* proviene de *cantus,* deverbal de *canere* (*canere*>*cantus*).

(1) **SINCRONÍA**

V	→	N
cantar		canto
trasladar		traslado
comprar		compra

DIACRONÍA

latín		**español**
canere, cantus,-ūs	---	canto
cantāre, cantātio	---	cantar
transferre, translātus,-ūs	---	traslado, trasladar

Para los diccionarios etimológicos resulta imprescindible la identificación de las familias léxicas (Campos, 2008). El proyecto MORFOGEN tiene como otro de sus objetivos dar soporte o cobertura al NDHE de la RAE en cuanto a su organización genética.

1.2. En el campo de la didáctica de lenguas, son muchos los especialistas que han constatado las ventajas de una enseñanza sistemática del vocabulario, para lo cual resultan muy provechosas todas las estrategias basadas en la elaboración de constelaciones o racimos de palabras, sean de tipo estrictamente semántico, sean de base morfológica. En las citas recogidas en (2-4) se reconoce la utilidad de identificar las relaciones morfosemánticas en el aprendizaje del vocabulario:

(2) The important principle behind the idea of a word family is that once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort (Bauer & Nation, 1993: 253).

(3) Il s'agit de faire travailler l'apprenant sur les "familles" morpho-sémantiques où un lexème produit des mots de différentes catégories grammaticales, avec différents effets de sens, par le jeu des préfixes et des suffixes. Des règles de dérivation relativement simples permettent de passer du verbe *réparer* aux dérivés *réparation* (nom d'action), *réparateur* (nom d'agent) à l'adjectif *réparable*, et à son antonyme *irréparable* ou du nom *courage* à l'adjectif *courageux* et à l'adverbe *courageusement* (Picoche, 1999).

(4) Los alumnos experimentan entusiasmo y sorpresa cuando se enteran de que *considerar* significa etimológicamente 'consultar con los astros'; *alumno* es 'el que ha de ser alimentado'; *feliz* es 'la que da de mamar'; *entusiasmarse* es 'meterse dentro de Dios' o 'tener dentro un Dios'; y *menopausia* es 'el descanso del mes'. (Campa *et al.,* 2008: *Introducción*).

[2] La datación está tomada del DCECH.

Las afirmaciones anteriores respaldan, por tanto, la tesis de que las relaciones morfológicas y las etimológicas resultan imprescindibles para comprender el significado y la evolución de las palabras. Para dar respuesta a esta necesidad, en el proyecto MORFOGEN se ha creado una aplicación web (bdme.usc.es) integrada por dos módulos: a) una base de datos de morfología del español (BDME) que, gracias a su reformulación digital, permite el trabajo colaborativo de varios usuarios simultáneos así como la realización de búsquedas de datos combinados de varios tipos (§5); b) una herramienta de visualización de familias léxicas.

En esta contribución se examinan los siguientes aspectos: en el §2 se analizan las diferentes maneras de concebir la noción de *familia léxica*; en el 3 se resumen las fases del proyecto MORFOGEN entre 2009 y 2012 y las modificaciones llevadas a cabo en ese período. El §4 se dedica a la descripción de la BDME en su configuración actual y a las búsquedas que permite. El §5 está destinado a mostrar dos formas de visualización de los datos. El §6 hace referencia a futuros trabajos y posibles desarrollos de la aplicación.

2. LA NOCIÓN DE *FAMILIA LÉXICA*

Una noción central para este proyecto es el concepto de *familia léxica de palabras* (*word family, famille de mots*). La etiqueta de familia hace referencia a relaciones de parentesco y el adjetivo *léxico* remite a un tipo de significado. Por este motivo, una base de datos léxica puede albergar diferentes agrupaciones de palabras. Al menos son tres las posibilidades que hemos identificado:

2.1. En las denominadas bases de datos léxicas, las palabras se agrupan por relaciones semánticas. Constituyen lo que también se conoce como "campos semánticos"[3]; una versión antigua de estas agrupaciones la encontramos en los diccionarios de ideas afines, y parcialmente en los de sinónimos y antónimos. Una aplicación web para el inglés que trabaja con asociaciones de significados y de conceptos es *VISUWORDS Online Graphical Dictionary*, donde las diferentes relaciones se marcan con diferentes trazos de líneas y colores, y cuya representación pretende imitar redes neuronales. Entre las relaciones que considera tienen cabida las siguientes: (i) es un tipo de; (ii) es un ejemplo de; (iii) es un miembro de; (iv) es una parte de; (v) es una sustancia de; (vi) es similar a; (vii) pertenece a; (viii) participio; (ix) atributo; (x) se opone a; (xi) grupo verbal; (xii) implica, etcétera.

Como puede deducirse fácilmente de la lista, se trata de relaciones muy heterogéneas, ya que se combinan asociaciones conceptuales (algunas próximas a la heteronimia, la sinonimia o la antonimia), con relaciones lógicas (implicación) y con conexiones de carácter gramatical (participio, grupo verbal). Por ello, los resultados obtenidos pueden ser muy útiles desde el punto de vista conceptual, pero ofrecen menos interés desde un punto de vista puramente lingüístico. Debemos admitir, no obstante, que la forma de plasmar gráficamente estas relaciones ha servido de inspiración a los informáticos que han diseñado nuestra herramienta de visualización.

[3] También puede consultarse el proyecto *Wordnet*, cuya idea central es la organización del léxico en campos semánticos (<http://wordnet.princeton.edu/>).

2.2. La base de datos francesa POLYMOTS se apoya en un concepto morfofonológico de familia[4]: a partir de una forma radical o palabra base, que se considera una unidad recurrente y que puede tener o no significado en la lengua, se agrupan las palabras derivadas. Sobre dicha forma radical se pueden crear palabras transparentes, como las que originan *bras* 'brazo', *fil* 'hijo' o *table* 'mesa', ya que combinan realización fónica y sentido, o bien palabras opacas, como *ferv–, duct–*, *oval–*, que no se asocian a un contenido en francés actual. La presencia de estos elementos se debe a que las autoras de esta base de datos, expertas en procesamiento del lenguaje natural, han constatado que una agrupación limitada a relaciones formales transparentes dejaba sin atender buena parte del vocabulario francés de origen culto (palabras como *conduire, conducteur*). Por este motivo, han agrupado bajo un radical opaco como *duct/du* un conjunto de 70 palabras que guardan alguna relación con la latina *ducere* (*duct–* es la forma de *perfectum* latino y *–du–* es la forma que aparece en algunos derivados franceses como *conduire*). El inconveniente de este análisis es que, sin atender a la diacronía, se propone segmentar en sincronía el verbo *conduire* del siguiente modo: prefijo *con–,* radical *–du–* y desinencia *–ire*[5].

2.3. El tercer punto de vista es el que realmente nos interesa. Se trata del concepto de *familia léxica*, que tiene en cuenta relaciones morfo-semánticas o relaciones derivativas. Ahora bien, existen al menos dos formas de concebir las relaciones morfo-semánticas.

2.3.1. Adoptando una perspectiva exclusivamente sincrónica, se considerarán parte de una misma familia léxica todas aquellas palabras que posean semejanza formal de la que deriva cierta proximidad significativa. Este enfoque tiene la ventaja de que permite abarcar términos pertenecientes a diferentes clases de palabras. Existe una aplicación web, desarrollada por el Grupo de Estructuras de Datos y Lingüística computacional de la Universidad de Las Palmas de Gran Canaria, que lleva a cabo esta tarea y representa las palabras del español en una especie de árbol, que enlaza la palabra cabeza con sus derivados (se denomina "detector automático de relaciones morfológicas derivativas en el léxico del español"). En palabras de los autores de esta aplicación (Santana *et al*., 2013):

> Como fruto del proceso formativo que ha sufrido una palabra, se le asocian sus relaciones morfoléxicas, se define el mecanismo por el cual se establecen y su regularidad en la formación. Se recopilan todos los procesos formativo-derivativos del español sin entrar en discusiones teóricas de uso, formalismos históricos y otros aspectos poco prácticos desde el punto de vista del procesamiento del lenguaje natural dirigido a la automatización de procesos lingüísticos, sin obviar, en casos puntuales, los procesos histórico-etimológicos de las palabras.

Hemos realizado una búsqueda de las relaciones morfoléxicas de la palabra *negro.* Además de la página que muestra el listado de sus relaciones derivati-

[4] *Vid.* Gala y Rey (2008).

[5] El proyecto fue diseñado para construir las familias de forma semiautomática, a partir de una lista de palabras extraídas del diccionario Larousse 2000 y del TLFi. Según declaran sus realizadoras, el objetivo inicial era fundamentalmente pedagógico: aprendizaje del vocabulario y de la ortografía del francés, pero en la actualidad se busca que tenga utilidad para la comunidad lingüística y para quienes trabajan en tratamiento automático de lenguas (Gala y Rey, 2008).

vas, también se puede obtener el grafo asociado. Esta aplicación ofrece varios resultados valiosos: indica la categoría léxica de la palabra (sustantivo, adjetivo, verbo, etc.), señala con colores los cambios de categoría en el proceso derivativo (el color rosa marca el cambio de adjetivo a verbo, el amarillo de adjetivo a sustantivo) y separa los afijos de cada palabra derivada. La familia que representa el árbol de las relaciones morfoléxicas de *negro* está formada por un total de 25 términos, con el siguiente reparto por categorías léxicas:

(5) S = greno, negrada, negralla, negrería, negrerío, negrismo, negritud, negrizal, negror, negrura = 10
A = negral, negrero, negrestino, negreta, negrillo, negrito, negrizco, negroide, negruzco, nigérrimo, seminegro = 11
V = negrear, negrecer, negregar, denegrir, ennegrecer = 5

Los autores de esta aplicación reconocen la necesidad de ampliar las relaciones morfo-léxicas más allá de las estrictamente morfosemánticas, incluso desde la óptica de los análisis automáticos. En palabras de Santana y otros (2013): "La potencialidad del tratamiento automático de la lengua quedaría cercenada si las relaciones que se establecen entre las palabras se restringieran a las puramente morfoléxicas. Se considera de interés ampliar la concepción de relación morfoléxica –relación morfoléxica extendida– en lo que atañe tanto a la raíz como a los afijos". Como consecuencia de ello, proponen extrapolar el concepto de afijo a elementos que "no siendo tales, actúan como si lo fueran". Ello implica introducir la perspectiva diacrónica en el análisis, al incluir casos como los de *lácteo, osificar*, que mantienen con sus respectivas bases *–leche, hueso–* relaciones semánticas codificadas en la morfología, como las de 'perteneciente o relativo a' y 'hacer o convertir', respectivamente. El problema es que también se incluyen como afijos segmentos puramente fonológicos (en palabras descendientes del árabe, segmentan el artículo *al–* o la terminación *–alle* en vocablos procedentes del francés o del catalán).

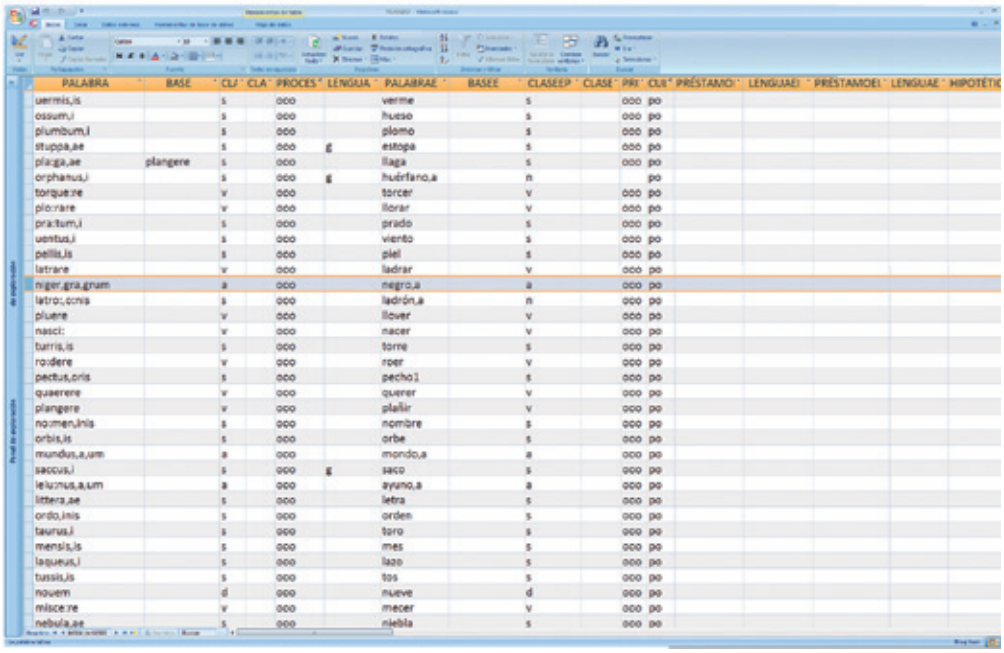
2.3.2. Adoptando una perspectiva diacrónica, la familia se contempla desde una concepción genética o etimológica y se concibe como un "conjunto de palabras agrupadas en torno a una raíz, étimo o palabra primitiva, como punto de arranque de las demás palabras emparentadas" (Pena y Campos Souto, 2009: 32).

La BDME responde a esta concepción, ya que pretende servir como estructura-soporte de un diccionario histórico. Desde una perspectiva histórica, la familia léxica es una agrupación de palabras más amplia o abarcadora que otra concebida desde el punto de vista sincrónico. Tiene que agrupar todas aquellas palabras que directa o indirectamente remiten a una misma raíz, sean estas variantes cultas, populares o préstamos e independientemente del mayor o menor grado de transparencia formal y semántica que plasmen en cuanto a la relación derivativa.

Como consecuencia de la aplicación de este enfoque, se pueden obtener series heterogéneas de derivación, donde se muestran las alternancias forma popular-forma culta (cf. la macrofamilia de *hacer*, que consta de, al menos, 650 voces, con la que ejemplifican Pena y Campos, 1999: §3). Los grados de parentesco son variados –algunos más evidentes o inmediatos y otros menos evidentes–, pero siempre se pueden establecer capas o estratos que permitan identificar subfamilias dentro de esas familias más amplias hasta llegar a series de derivación más homogéneas formal y semánticamente.

3. FASES DEL PROYECTO

La primera fase del proyecto consistió en la normalización y adaptación de una base de datos previa en formato ACCESS (con información sobre latín, español y préstamos). Esta base acumulaba toda la información de las diferentes lenguas en una tabla única y codificaba con letras los distintos valores asignados en cada uno de los campos a cada nuevo vocablo analizado. La figura 1 deja ver algunos campos de la base y la codificación alfabética de las propiedades de las palabras en cada campo.



PALABRA	BASE	CLI	CLA	PROCES	LENGUA	PALABRAE	BASEE	CLASEEP	CLASE	PRI	CUL	PRÉSTAMO	LENGUAE	PRÉSTAMOEI	LENGUAE	HIPOTÉTIC
uermis,is		s		ooo		verme		s		ooo	po					
ossum,i		s		ooo		hueso		s		ooo	po					
plumbum,i		s		ooo		plomo		s		ooo	po					
stuppa,ae		s		ooo	g	estopa		s		ooo	po					
pla:ga,ae	plangere	s		ooo		llaga		s		ooo	po					
orphanus,i		s		ooo	g	huérfano,a		n			po					
torque:re		v		ooo		torcer		v		ooo	po					
plo:rare		v		ooo		llorar		v		ooo	po					
pra:tum,i		s		ooo		prado		s		ooo	po					
uentus,i		s		ooo		viento		s		ooo	po					
pellis,is		s		ooo		piel		s		ooo	po					
latrare		v		ooo		ladrar		v		ooo	po					
niger,gra,grum		a		ooo		negro,a		a		ooo	po					
latro:,onis		s		ooo		ladrón,a		n		ooo	po					
pluere		v		ooo		llover		v		ooo	po					
nasci:		v		ooo		nacer		v		ooo	po					
turris,is		s		ooo		torre		s		ooo	po					
ro:dere		v		ooo		roer		v		ooo	po					
pectus,oris		s		ooo		pecho1		s		ooo	po					
quaerere		v		ooo		querer		v		ooo	po					
plangere		v		ooo		plañir		v		ooo	po					
no:men,ĭnis		s		ooo		nombre		s		ooo	po					
orbis,is		s		ooo		orbe		s		ooo	po					
mundus,a,um		a		ooo		mondo,a		a		ooo	po					
saccus,i		s		ooo	g	saco		s		ooo	po					
ie:iunus,a,um		a		ooo		ayuno,a		a		ooo	po					
littera,ae		s		ooo		letra		s		ooo	po					
ordo,inis		s		ooo		orden		s		ooo	po					
taurus,i		s		ooo		toro		s		ooo	po					
mensis,is		s		ooo		mes		s		ooo	po					
laqueus,i		s		ooo		lazo		s		ooo	po					
tussis,is		s		ooo		tos		s		ooo	po					
nouem		d		ooo		nueve		d		ooo	po					
misce:re		v		ooo		mecer		v		ooo	po					
nebula,ae		s		ooo		niebla		s		ooo	po					

Fig. 1. La BDME con su formato en Access.

En esta base los campos 2 a 9 se referían a palabras latinas. Los demás a palabras del español actual. Estaba construida sobre una matriz de rasgos que proporcionaba información completa sobre los aspectos formales, funcionales y semánticos de la palabra: (i) Clase y subclase de la palabra base y de la derivada; (ii) Tipo formal de la palabra base y de la derivada: popular/culto; (iii) Tema de palabra base inexistente como palabra; (iv) Estructura mórfica de la palabra; (v) Procedimientos morfológicos utilizados en la creación de la palabra derivada; (vi) Serie de derivación de la que forman parte la palabra base y la palabra derivada; (vii) Palabra de creación autóctona, frente a palabra incorporada como préstamo. Distinción entre préstamo inmediato y mediato; (viii) Marcas de uso. Cada campo llevaba asociado un conjunto de marcas (símbolos en forma de letras) que permitían llevar a cabo la clasificación de cada una de las palabras de la base.

La remodelación de esta base de datos con vistas a su conversión digital se realizó en dos etapas. En la primera, se configuraron 12 tablas, las principales para el español y el latín, más otra para griego y una específica para temas greco-latinos (tipo *arboricida, adenopatía*, etc.), así como tantas tablas como lenguas intervienen, de forma recurrente, en la incorporación de préstamos. Cada tabla contenía el mismo tipo de información para todas las lenguas consideradas, a saber portugués, gallego, occitano, catalán, francés, italiano, inglés. Existía, por último, una tabla específica para otras lenguas, con menos propiedades. Al desarrollar la aplicación se constataron dos problemas: a) La tabla para otras lenguas se había simplificado en exceso y no permitía el análisis de préstamos de lenguas como el árabe o el alemán (términos provenientes del árabe antiguo o del alto alemán, por ejemplo); b) La tabla para temas grecolatinos, no se había vinculado a lenguas específicas, por tratarse de temas de palabras inexistentes. En el año 2010 se introdujeron algunos temas a partir del Diccionario Vox, con vistas a una posterior asignación como bases de palabras españolas. Al comprobar que el vocabulario científico se incorpora a través del francés y del inglés, se vio la necesidad de vincular muchos temas a estas dos lenguas; además, en las propias lenguas griega y latina hay temas de palabras inexistentes como palabras simples que sólo aparecen en cuanto constituyentes de palabras afijadas y compuestas.

Estos problemas obligaron a reconsiderar el proceso de almacenamiento de datos y sus relaciones. Desde marzo de 2012 existe una tabla única, en la que las lenguas se seleccionan mediante un menú desplegable, con el siguiente orden: a) las más usadas en el proyecto; b) las restantes por orden alfabético. El formulario de introducción de datos también se modificó, para dar cabida a campos como SIGNIFICADO, TIPOS DE AFIJOS, EJEMPLOS, DATACIÓN, etc. Al analizar una palabra, el investigador debe seleccionar entre clasificarla como subclase de palabra (adjetivo, verbo, etc.) o como tema.

Se ha dedicado especial atención en el proyecto a los siguientes aspectos: a) relación de lenguas posibles como fuentes de préstamo para el español; b) relación de sufijos y prefijos, establecida siguiendo pautas diacrónicas; c) significados morfológicos (una lista cerrada de posibles significados esquemáticos de las palabras derivadas); d) marcas de uso, que se han adaptado a las diferentes lenguas. En el caso del español, se han trasladado algunas marcas del DCECH y también se ha buscado que las marcas geográficas armonicen con las del NDHE. Se distinguen dos marcas de tipo histórico, una que sigue las convenciones del DRAE (antigua, desusada, poco usada) y otra que sigue las agrupaciones del NDHE.

4. CONFIGURACIÓN ACTUAL DE LA BDME

La BDME es una base de datos relacional, es decir, los datos se almacenan en varias tablas conectadas entre sí para permitir la interconexión de los datos recogidos en ellas. En concreto, la BDME está compuesta por un total de trece tablas que se organizan en torno a una principal, la tabla TÉRMINO. Esta tabla contiene veinte campos que permiten recoger distinta información relativa a cada una de las palabras que se almacenan en la base de datos: el idioma, el lema, la clase de palabra, los sinónimos, si tiene estatus de palabra o se considera un tema de palabra inexistente, si se trata de una palabra hipotética,

las fechas de su primera y última documentación o ejemplos de su uso, entre otras propiedades.

Algunos de estos campos están relacionados con información albergada en otras tablas. Es el caso, por ejemplo, del idioma. A cada una de las lenguas utilizadas en la base de datos está asociado un color con el que se representarán en el grafo y la visualización lineal; también se indica si se trata de una lengua que aparece con frecuencia en la base de datos o no, un parámetro que se utiliza para ordenar las lenguas en el menú desplegable del formulario web empleado para introducir los datos. Es decir, cuando se introduce el idioma de una palabra en la tabla TÉRMINO, automáticamente la palabra en cuestión pasa a estar relacionada también con la información almacenada acerca de esa lengua en la tabla IDIOMA.

La información relativa a los significados morfológicos, las marcas de uso y las bases de derivación se almacenan también en tablas independientes, aunque vinculadas a TÉRMINO de un modo similar al de la tabla IDIOMA. Por ejemplo, cuando una palabra actúa como base de derivación de otra, pasa a estar almacenada también en la tabla BASE, junto con el proceso mediante el cual se ha formado la palabra. En aquellos casos en los que el proceso de formación de la palabra es la prefijación, la sufijación o ambas, entran en juego otras dos tablas: PREFIJO y SUFIJO, que contienen información acerca de cada afijo, como las variantes o algunos ejemplos de palabras en las que aparecen.

También las tablas que contienen las marcas de uso y los significados morfológicos tienen asociadas a su vez otras tablas, que permiten jerarquizar la información. Las marcas de uso, por ejemplo, se organizan en cuatro niveles para el español y cada tipo de marca admite subdivisiones. Esto permite dar cuenta de la inclusión de, por ejemplo, Bolivia, Ecuador y Perú como países del área andina, variedad del español de América que es una marca de tipo geográfico.

Por otra parte, se han diseñado dos buscadores de términos. El primero, básico, permite búsquedas alfabéticas por inicio, medio o fin de palabra. Este buscador fue muy importante en la primera fase de remodelación de la base, ya que se hizo una importación de datos desde la BDME en formato ACCESS que, por diferentes motivos, no fue todo lo exitosa que debiera. Con buscadores de este tipo fue posible saber qué palabras no se habían introducido en la importación y cuáles sí.

El segundo buscador, denominado avanzado, permite búsquedas específicas combinando propiedades diferentes y lenguas diferentes. En la figura 2 se muestra la búsqueda de palabras del IDIOMA español, de la categoría SUSTANTIVO, que tengan origen en palabras latinas [ORIGEN, IDIOMA, Latín] cuya base sea un Verbo.

El resultado de la búsqueda se presenta en la figura 3. La pantalla informa del número de resultados (casi 1.900 términos), que se muestran ordenados alfabéticamente y con las propiedades de la tabla que el usuario ha seleccionado previamente: TÉRMINO, ENTRADA, ORIGEN Y BASE. Desde esta pantalla (o las sucesivas) se puede acceder tanto a la ficha de cada término como al grafo de la familia léxica que le corresponde.

Fig. 2. Buscador avanzado de términos en la BDME.

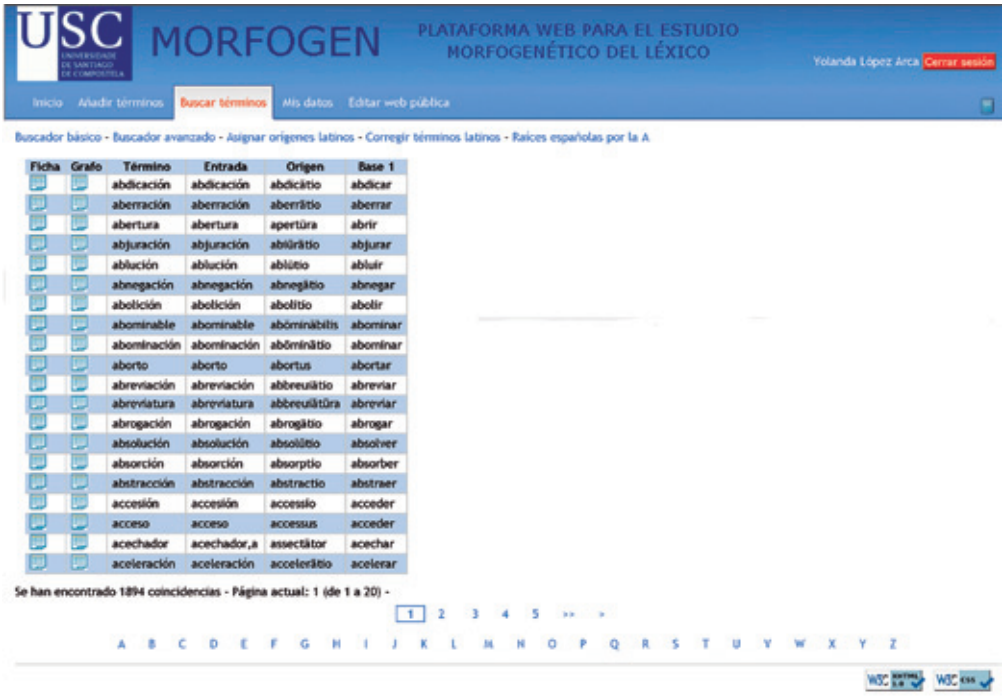


Fig. 3. Pantalla de resultados del buscador avanzado.

5. VISUALIZACIÓN DE LOS DATOS

5.1. En forma dinámica

Se ha diseñado una herramienta de visualización que permite la búsqueda de una familia léxica a partir de un término almacenado en la base. El programa realiza una consulta a la base de datos y representa la familia léxica mediante un grafo o árbol con sus nudos, teniendo en cuenta dos tipos de relaciones: a) la que vincula un término con su base, es decir, aquella palabra de la misma lengua de la que deriva; b) la que vincula un término con su origen, es decir, aquella palabra de otra lengua de la que es continuación, incorporación o adaptación. La herramienta distingue las lenguas con colores y tiene en cuenta las clases de palabras. En la figura 4 aparece la pantalla de la familia léxica del adjetivo *negro*[6], en su forma abreviada, agrupando los derivados según su categoría (S = sustantivo, A = Adjetivo, V = Verbo, O = Otras). Esta agrupación en clases aparece cuando el número de derivados es superior a 8 palabras.

Los términos ocultos pueden desplegarse haciendo clic con el ratón sobre cada uno de estos círculos. Es posible volver a ocultarlos mediante un segundo clic. También pueden mostrarse u ocultarse la base o el origen de un término haciendo clic con el botón derecho sobre él. Los nudos correspondientes a cada

[6] En la BDME se distingue entre el adjetivo *negro,-a* y su recategorización como sustantivo.

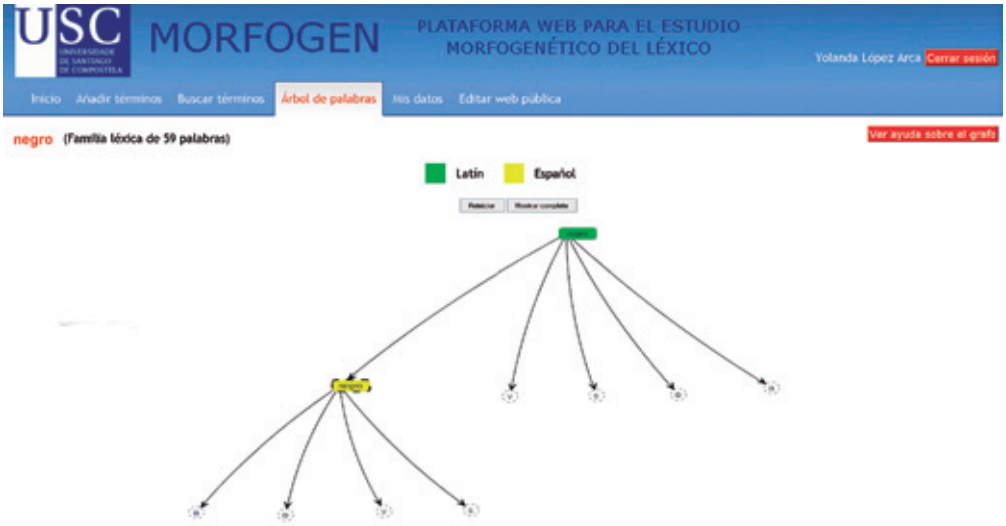


Fig. 4. Pantalla de la familia léxica de negro *presentada de forma abreviada.*

término pueden moverse para facilitar la visualización. Haciendo clic sobre ellos mientras se mantiene pulsada la tecla Ctrl se accede a la ficha completa en la base de datos.

En la figura 5 se ofrece la familia al completo, tanto en sentido descendente (palabras creadas en español sobre el adjetivo) como en sentido ascendente (palabras latinas originarias de las españolas). Un aspecto que necesita ser mejorado es la presentación de esta segunda imagen: los árboles de familias numerosas se muestran con ramas amontonadas y desordenadas, si bien el usuario puede mover las ramas y recolocarlas para que se visualicen mejor.

Por tanto, la representación en forma de grafo permite explorar la familia léxica de manera dinámica, profundizando en una determinada serie de derivación, eligiendo los elementos que interesen o prescindiendo de la información que no resulte interesante en una determinada consulta. Los elementos de cada lengua se representan con un determinado color que permite distinguirlos de los términos de lenguas distintas que estén implicados en la formación de la familia léxica. Las flechas que unen los términos señalan las relaciones que existen (i) entre las bases de derivación y sus derivados, (ii) entre el origen de una determinada palabra y su resultado y (iii) entre las palabras que sirven de préstamos y sus correspondientes resultados en otras lenguas.

5.2. En forma estática

La visualización lineal permite explorar una familia léxica de manera estática (Fig. 6). Para una lengua como el español reproduce de forma transparente la secuenciación de las relaciones derivativas (como N > V > A). Esta representación resulta especialmente útil para comprobar de manera gráfica qué palabras se han perdido en la evolución del latín al español y cuáles son creaciones hispánicas. La representación parte del término que da origen a la familia léxica, que carece tanto de base de derivación como de origen. En cada línea se muestran una a

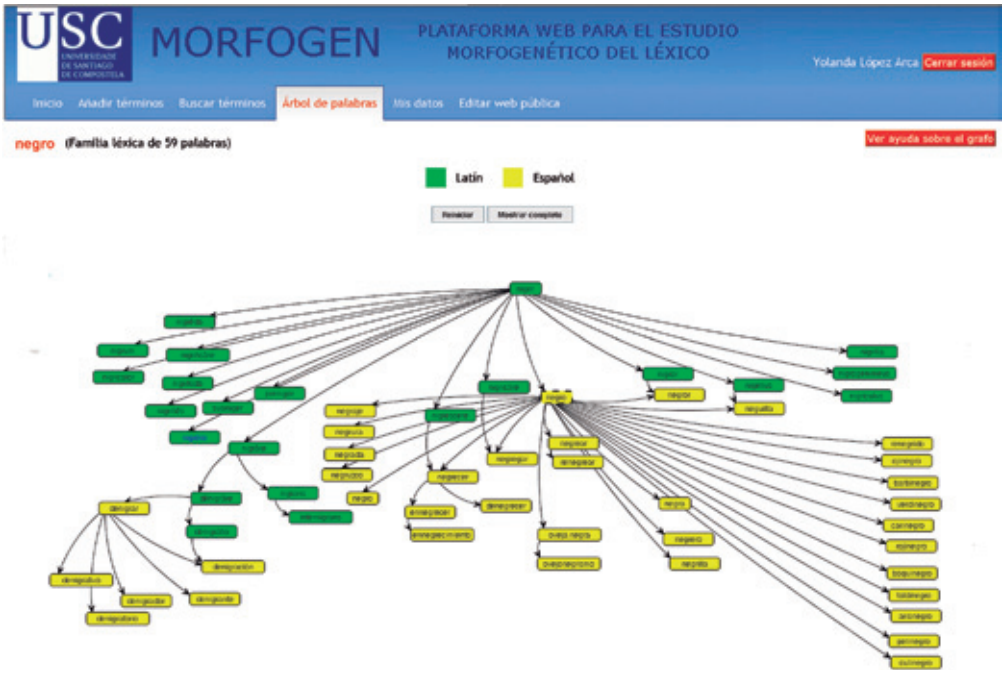


Fig. 5. Pantalla de la familia léxica del adjetivo negro*, en relación con el latín* niger *y sus derivados. Los resultados han sido ordenados por el usuario.*

una las series de derivación que forman la familia. De izquierda a derecha se representan las relaciones entre las bases de derivación y sus derivados.

En la representación pueden alternarse series de derivación en más de una lengua. Los elementos de cada lengua se identifican con un mismo color, que coincide con el del grafo. Si un término es el antecedente de otro término en una lengua distinta, su correlato aparece en la línea inferior. Las relaciones por la vía de origen se representan, pues, en el eje vertical, de arriba hacia abajo (cf. *niger-negro*, *nigror-negror, nigellus-neguilla, nigrescere-negrecer*, etcétera).

Las series de derivación en la lengua que da lugar a la familia léxica se ordenan según la clase de palabras a la que pertenezca el primer derivado de la serie. En primer lugar aparecen los sustantivos (*negror, negrilla*), seguidos de los adjetivos y los verbos (*negrecer, denegrecer, ennegrecer…*). Por último aparecen las palabras analizadas como adjetivos y sustantivos, adjetivos y participios, etc. En los casos en los que hay más de un derivado de la misma clase se ordenan alfabéticamente. En último lugar, en la parte inferior de la representación, aparecen las series de derivación cuyo primer término derivado se crea en español sin antecedente latino. Los criterios de ordenación se mantienen: primero se clasifican según la clase de palabras, como se indicó anteriormente, y en segundo lugar por orden alfabético (*negrada, negraje, negrura…*) Las palabras compuestas se incluyen también en este último grupo (*carinegro, faldinegro…*).

La representación en paralelo de las series de derivación latina y española permite observar fácilmente qué palabras se han perdido en la evolución del latín al español, ya que los huecos que ocuparían esas palabras se corresponden con

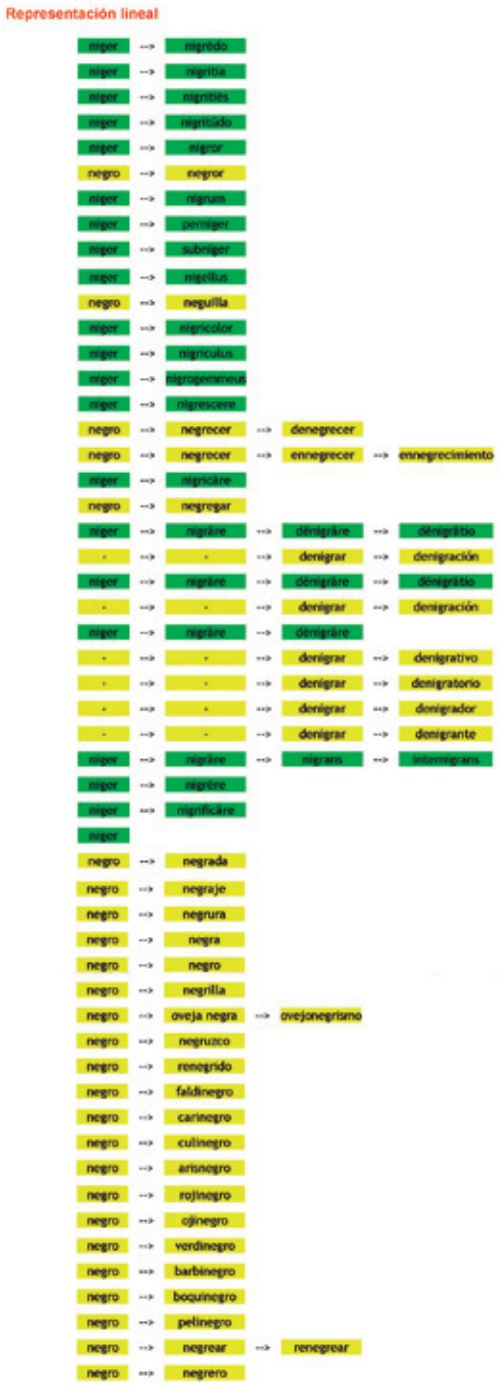


Fig. 6. Representación lineal parcial de la familia del adjetivo negro *(lo que permite ver una pantalla).*

blancos en las series derivativas. Al mismo tiempo, resulta fácil explorar cuáles son las creaciones hispánicas que carecen de antecedente latino buscando las casillas de color amarillo que ocupan un lugar más a la derecha que la última casilla verde de la fila superior.

6. PERSPECTIVAS DE FUTURO

En los próximos años el equipo del proyecto se propone revisar los datos actualmente almacenados para completar el análisis con las propiedades añadidas (significados morfológicos, sufijos y prefijos, marcas de uso, datación). El vocabulario culto, construido sobre temas grecolatinos, está pendiente de análisis, debido a la remodelación del formulario. La parte etimológica de la base, propiedad ORIGEN, es susceptible de revisión y están sin completar los préstamos de otras lenguas actuales y las palabras formadas sobre temas de palabras inexistentes como palabras simples, empezando por el griego y el latín, y continuando, por este orden, con el francés, inglés y español. Se pretende, además, aumentar el número de palabras analizadas, completando en lo posible las familias léxicas.

El segundo componente de la aplicación web, la herramienta de visualización, también debe ser mejorada en las siguientes facetas: (a) evitar la acumulación de información cuando las familias léxicas son extensas, para lo cual se ha estado ensayando con el plegado y desplegado de las ramas, con el movimiento de los grafos y con la agrupación de los nudos por categorías léxicas; (b) mejorar la información asociada a cada término en los nudos de los grafos; creemos que resulta especialmente relevante añadir pequeñas glosas con el significado de las palabras y, pensando en su explotación para la enseñanza del español, añadir los equivalentes en inglés de cada término; (c) utilizar colores en las flechas para distinguir relaciones derivativas, especialmente necesarios cuando el grafo se despliega en su totalidad; (d) asociar los grafos a búsquedas específicas que permitan filtrar la información que resulte de interés para el usuario.

Una vez que la plataforma esté accesible, en función del interés que suscite, tiene una configuración que permitiría reproducir el trabajo llevado a cabo con el español para otras lenguas románicas.

BIBLIOGRAFÍA

Bauer, Laurie y Nation, Paul, "Word families", *International Journal of Lexicography,* 6/4 (2003), pp. 253-279.

Campa, Hermenegildo de *et al.*, *Diccionario español escolar de familias etimológicas*, Granada, Edición experimental, 2008, <http://es.scribd.com/doc/106187096/Diccionario-espanol-escolar-de-familias-etimologicas> [01/07/2013].

Campos, Mar, "Hacia la ordenación morfológica del NDHE: primer esbozo", *Verba,* 34 (2007), pp. 125-155.

DCECH, Corominas, Joan y Pascual, José Antonio, *Diccionario crítico etimológico castellano e hispánico*, Madrid, Gredos, 1980-1991.

Gala, Nuria y Rey, Véronique, "Polymots: une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques", *Traitement Automatique des Langues Naturelles, TALN 2008*, Avignon, 2008, <http://www.florianboudin.org/taln_archives/TALN/TALN-2008/actes/taln-2008-court-010.pdf> [01/07/2013].

NDHE, Fundación Instituto de Investigación Rafael Lapesa para el Nuevo diccionario histórico del español, <http://www.frl.es/Paginas/default.aspx>.

Pena, Jesús, "Alteraciones de la serie derivativa verbo-nombre deverbal en español. Análisis genético", *Zeitschrift für romanische Philologie,* 128/2 (2012), pp. 319-349.

Pena, Jesús y Campos, Mar, "Propuesta metodológica para el establecimiento de familias léxicas: el caso de *hacer*", *Cuadernos del Instituto Historia de la lengua*, 2 (2009), pp. 21-51.

Picoche, Jacqueline, "Dialogue autour de l'enseignement du vocabulaire", *Études de Linguistique Appliquée,* 116 (1999), pp. 421-434, <http://jpicochelinguistique.free.fr/ENSEIGNEMENT/articlesdefond.html>.

POLYMOTS, <http://polymots.lif.univ-mrs.fr/v2/> [01/07/2013].

Relaciones morfoléxicas, O. Santana *et al.,* <http://www.gedlc.ulpgc.es/investigacion/scogeme02/relmorfo.htm> [01/07/2013].

VISUWORDS *Online Graphical Dictionary*, <http://www.visuwords.com/> [01/07/2013].

WORDNET, <http://wordnet.princeton.edu> [01/07/2013].

RESUMEN

El objetivo de esta contribución es presentar las características de una aplicación web destinada a hacer visibles en español las relaciones de parentesco entre las palabras de una misma familia léxica (<bdme.usc.es>). La aplicación contiene dos módulos: a) una base de datos de morfología del español y b) una herramienta de visualización. En este trabajo se mostrarán diferentes concepciones de la noción de familia léxica y también algunos procedimientos empleados en otros proyectos para dar cuenta de relaciones morfoléxicas. Se mostrará la conveniencia de aplicar un enfoque histórico a las familias léxicas y de trabajar con dos tipos de relaciones de parentesco: (i) la que asocia una palabra con su origen y (ii) la que vincula una palabra con su base léxica. Se comentarán algunos problemas surgidos durante el proceso de conversión de una base de datos monousuario en otra de acceso compartido. Se explicarán las principales características técnicas de la base de datos y se mostrará el diseño de dos formas de representación: en diagramas arbóreos y en forma lineal. Con la primera se pretende ofrecer una imagen global de cada familia léxica y con la segunda se busca reflejar la progresión de los procesos derivativos en las series de una misma lengua.

Palabras clave: Familia léxica, morfología histórica, etimología.

ABSTRACT

This paper shows the characteristics of a web application designed to represent the family relationships between components of a word family in Spanish (<bdme.usc.es>). The application consists of two modules: a) a morphological database of Spanish and b) a visualization tool. Different formulations of the concept of 'word family' will be shown, as well as some procedures that have been used in other projects in order to account for morpho-lexical relationships. The convenience of a historical approach to word families will be discussed, as well as the appropriateness of working with two types of family relationships: (i) the one associating a word with its origin and (ii) the one associating the word with its lexical base. Some issues that emerged during the process of conversion from a single-user to a multiple-access database will be discussed. The main features of the database will be explained. The design of two representational forms (tree and linear diagrams) will be shown: the former aims to produce a global representation of a word family while the latter seeks to represent the progression of derivational processes in the series of a given language.

Keywords: Word family, historical morphology, etymology.