Manuel Casado Velarde[1] / Fernando Sánchez León[2]

# Notes on Real Academia Española's tools and resources

### Resumen / Abstract

Este documento esboza las herramientas y recursos más sobresalientes que está desarrollando la Real Academia Española para su inclusión en el nuevo portal web de la Institución. El proyecto del nuevo portal web, que verá la luz en 2011, pretende ser el eje vertebrador que permita dar a conocer la profunda renovación tecnológica que se ha realizado en el seno de la Academia. Por tanto, con alguna excepción, no se mencionan, de forma intencionada, las versiones actuales de estos recursos que se encuentran accesibles en las páginas de la Academia (www.rae.es), por considerar que son suficientemente conocidos para la comunidad hispanohablante.

## 1. Introduction

This document outlines the main tools and resources that the Spanish Royal Academy (Real Academia Española, RAE) is developing to include on its new web site, which is scheduled for publication in 2011. This webpage will demonstrate the profound technological renovation undertaken at the heart of the Academy. Therefore, with occasional exceptions, we deliberately do not mention here the present version of the resources now available on the Academy site (www.rae.es), since they are well known, at least to Spanish speaking scholars.

## 2. Basic aspects

Some years ago the RAE engaged itself in the creation of new linguistic resources that may supplement those which were historically developed in order to prepare its linguistic works. In this sense, the Academy owns at the moment a Spanish Data Bank constituted by three reference corpora: CORDE (*Corpus Diacrónico del Español*, 'Spanish Diachronic Corpus'), CREA (*Corpus de Referencia del Español Actual*, 'Present Day Spanish Reference Corpus'), and CORPES (*Corpus del Español del Siglo XXI*, '21st Century Spanish Corpus'), supplemented by several specialized corpora: CDH (*Corpus del Diccionario Histórico*, 'Historical Dictionary Corpus'), *Corpus Escolar* ('School Corpus') and CCT (*Corpus Científico-Técnico*, 'Scientific-Technical Corpus'), among others.

The RAE has taken care to develop its own linguistic technology which allows for both Spanish Data Bank linguistic tagging and for easier consultation of the rest of its linguistic resources, such as dictionaries of several kinds, the academic grammar, lexical lists index cards, or bibliographical data bases on Spanish vocabulary. Investment in linguistic technology began modestly some ten years ago, with the use of programs and tools already available in the public domain as a result of European projects like MULTEXT or CRATER. At the moment, the Technology Department of the Academy itself draws up text segmentation and morphological analysis and disambiguation programs, as well as the associated lexicons to these automatic tasks. Similarly, the RAE has other resources

[1] Correspondent Member, Spanish Royal Academy (RAE).
[2] Head of Technology Research, Center of Studies, Spanish Royal Academy (RAE).

linked to the diachronic analysis of Spanish that therefore allow the tagging of Spanish corpora throughout time. Finally, there are lexical resources taken from dictionaries of Latin American Spanish. They will enable the tagging of texts belonging to the different American varieties of Spanish.

The digitalization of the RAE's most valuable repositories is a priority for this institution. In fact, the Academy has already started to lay the foundations of a digital library containing the most important works from its archives; it has begun the digitalization of its lexical and lexicographical files. At present, there already exists a digital version of the Academy's *General File*, which includes all quotes appearing in the first edition of the Academic Dictionary, as well as the references accumulated by RAE's members and occasional collaborators in order to elaborate successive editions of the *Spanish Language Dictionary*. The total amount of index cards now approaches 12 million.

The interoperability between different resources is a short to medium term aim. The RAE is working on the idea of a "unified window" (*ventana única*) access to all those resources.

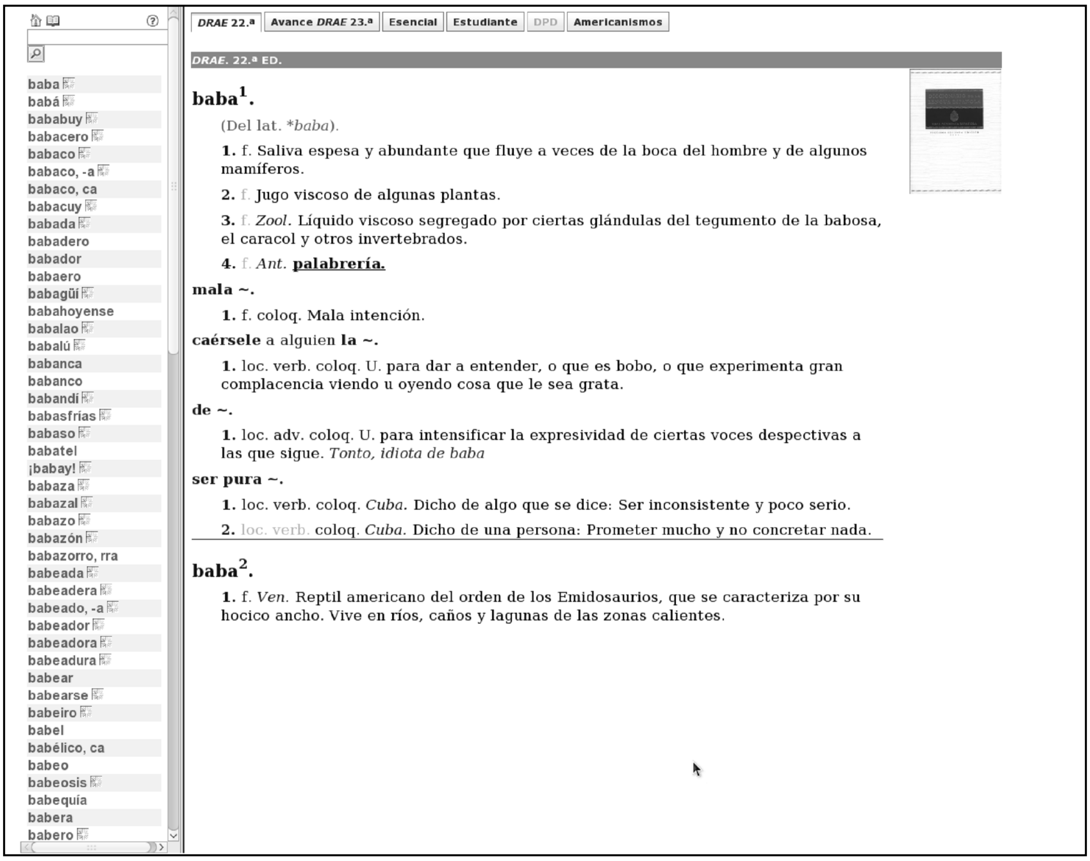
## 3. Applications

### 3.1 Lexicographic applications

Only two interfaces have been designed to consult the academic dictionaries: one for the synchronic dictionaries, and the other for the American Spanish lexicons. Along with these two interfaces, the application to consult the *Nuevo Tesoro Lexicográfico de la Lengua Española* (NTLLE, *New Lexicographic Thesaurus of Spanish Language*) is also preserved, until the future integration of all the academic dictionaries in a search only window.

The following three pictures provide some examples of these two interfaces. Figure 1 exhibits a unified search within the latest academic dictionaries. Consultable dictionaries in this application are, besides the DRAE, a preview of the 23rd edition of the aforementioned dictionary, the *Essential Dictionary* (*Diccionario esencial*), the *Student's Dictionary* (*Diccionario del estudiante*), the *Panhispanic doubts dictionary* (*Diccionario panhispánico de dudas*), and the *American Spanish Dictionary* (*Diccionario de americanismos*). Results appear on the left in a word list (all the headings beginning with *bab*-). By clicking on a word (in this case, *baba*, 'dribble') the window shows each article in the RAE's Spanish Dictionary (DRAE) in its current edition (the 22nd) . At the same time, a series of tabs to other dictionaries can also be seen, which are active only if the word is included in any of those particular dictionaries. Finally, in this case, this word has two homographs (*baba*$^1$ and *baba*$^2$), which are displayed on the same screen.

Figure 2 reveals the same word (*baba*) with its American meanings, as they have been listed in the *American Spanish Dictionary*. Access to this information is as simple as clicking on the appropriate tab on the application screen, provided that it is not dimmed (in grey), which would mean that the word is not included in that dictionary. This application greatly simplifies the consultation of dictionaries thanks to its ergonomics, which save users much search time.



**baba**[1]**.**

(Del lat. **baba*).

**1.** f. Saliva espesa y abundante que fluye a veces de la boca del hombre y de algunos mamíferos.

**2.** f. Jugo viscoso de algunas plantas.

**3.** f. *Zool.* Líquido viscoso segregado por ciertas glándulas del tegumento de la babosa, el caracol y otros invertebrados.

**4.** f. *Ant.* **palabrería.**

**mala ~.**

**1.** f. coloq. Mala intención.

**caérsele** a alguien **la ~.**

**1.** loc. verb. coloq. U. para dar a entender, o que es bobo, o que experimenta gran complacencia viendo u oyendo cosa que le sea grata.

**de ~.**

**1.** loc. adv. coloq. U. para intensificar la expresividad de ciertas voces despectivas a las que sigue. *Tonto, idiota de baba*

**ser pura ~.**

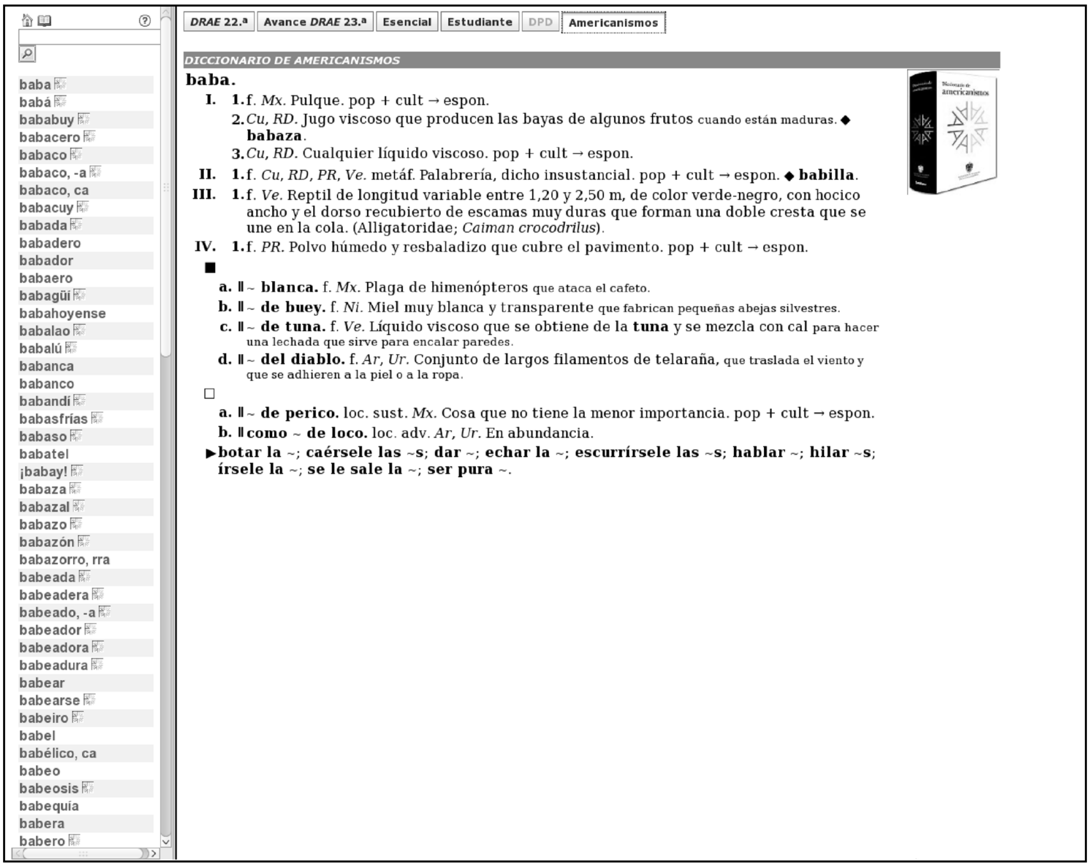
**1.** loc. verb. coloq. *Cuba.* Dicho de algo que se dice: Ser inconsistente y poco serio.

**2.** loc. verb. coloq. *Cuba.* Dicho de una persona: Prometer mucho y no concretar nada.

**baba**[2]**.**

**1.** f. *Ven.* Reptil americano del orden de los Emidosaurios, que se caracteriza por su hocico ancho. Vive en ríos, caños y lagunas de las zonas calientes.

Figure 1: Unified search within the latest academic dictionaries



**baba.**

**I.** **1.** f. *Mx.* Pulque. pop + cult → espon.
**2.** *Cu, RD.* Jugo viscoso que producen las bayas de algunos frutos cuando están maduras. ◆ **babaza**.
**3.** *Cu, RD.* Cualquier líquido viscoso. pop + cult → espon.

**II.** **1.** f. *Cu, RD, PR, Ve.* metáf. Palabrería, dicho insustancial. pop + cult → espon. ◆ **babilla**.

**III.** **1.** f. *Ve.* Reptil de longitud variable entre 1,20 y 2,50 m, de color verde-negro, con hocico ancho y el dorso recubierto de escamas muy duras que forman una doble cresta que se une en la cola. (Alligatoridae; *Caiman crocodrilus*).

**IV.** **1.** f. *PR.* Polvo húmedo y resbaladizo que cubre el pavimento. pop + cult → espon.

■

**a.** ‖ ~ **blanca.** f. *Mx.* Plaga de himenópteros que ataca el cafeto.

**b.** ‖ ~ **de buey.** f. *Ni.* Miel muy blanca y transparente que fabrican pequeñas abejas silvestres.

**c.** ‖ ~ **de tuna.** f. *Ve.* Líquido viscoso que se obtiene de la **tuna** y se mezcla con cal para hacer una lechada que sirve para encalar paredes.

**d.** ‖ ~ **del diablo.** f. *Ar, Ur.* Conjunto de largos filamentos de telaraña, que traslada el viento y que se adhieren a la piel o a la ropa.

□

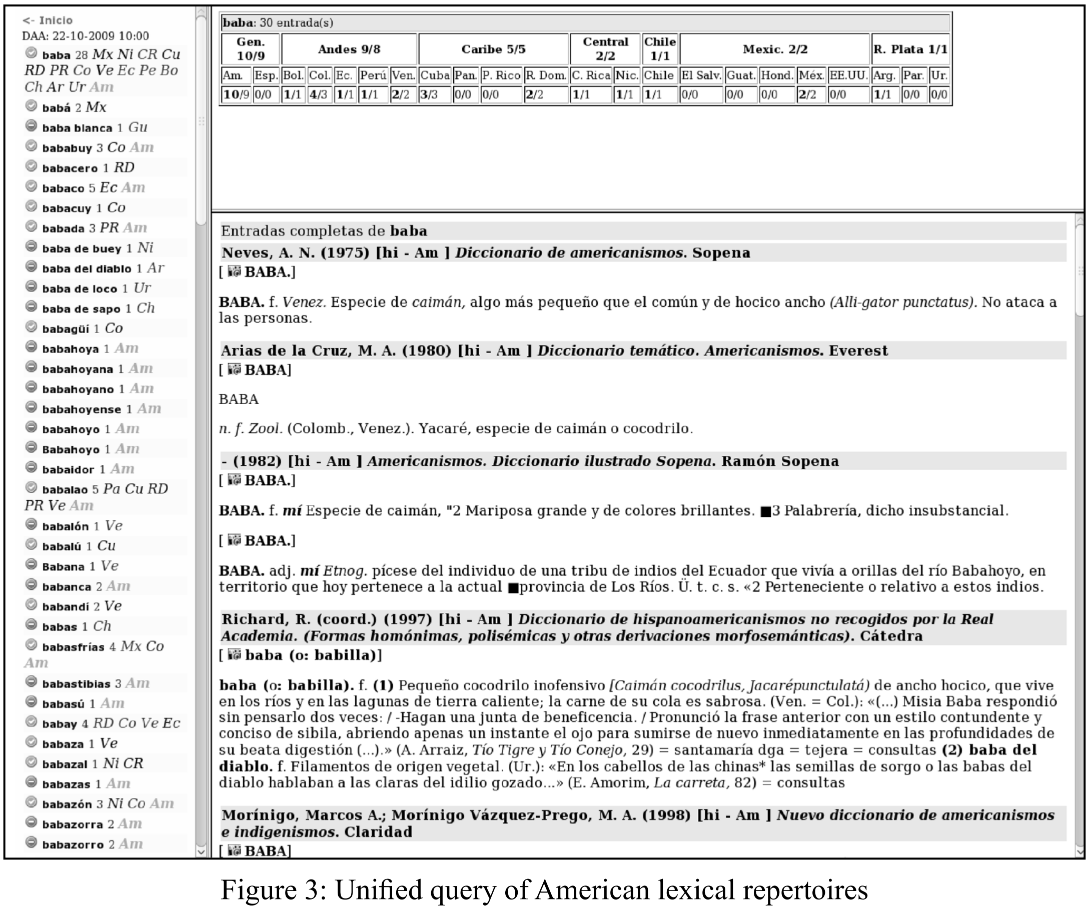
**a.** ‖ ~ **de perico.** loc. sust. *Mx.* Cosa que no tiene la menor importancia. pop + cult → espon.

**b.** ‖ **como ~ de loco.** loc. adv. *Ar, Ur.* En abundancia.

▶ **botar la** ~; **caérsele las** ~**s**; **dar** ~; **echar la** ~; **escurrírsele las** ~**s**; **hablar** ~; **hilar** ~**s**; **írsele la** ~; **se le sale la** ~; **ser pura** ~.

Figure 2: Meanings of *baba* in the *American Spanish Dictionary*

During the preparation of the *American Spanish Dictionary*, an application has been designed, with the same philosophy in mind, allowing the running of a unified query of over 120 American lexical repertoires. This application, called ARU (meaning 'word' and 'dictionary' in Aymara) shows the contents of these dictionaries in text format (see Figure 3). In this regard, all dictionaries that do not come from digital sources have been digitalized and have gone through an optical character recognition (OCR) reader. Only the nomenclature of this process outcome has been revised manually, using the double cross-validation method. At all times, users will be able to access the original image (the scanned page) by clicking on the icon that appears next to each lemma. A table displays the distribution of each word (in this case, *baba*, once again) in every dictionary depending on the language area in which each word is located. Finally, the wordlist (the query is conducted again on the word *bab**) exhibits, next to each heading, an icon indicating its presence or absence in the *American Spanish Dictionary*. Next to each heading there is a list of diatopic markings reflecting the overlap between the *American Spanish Dictionary* and the different sources available: black listed countries reflect matches between these two sets; blue diatopic markings indicate that, although that word is included in the *American Spanish Dictionary*, it does not appear in any of the ARU dictionaries; and the red ones represent the reverse situation. The mark *Am*, from general American Spanish dictionaries, it is not used in the *American Spanish Dictionary* of the Association of Academies; for this reason, it appears in a different colour, orange (in this case).



<- Inicio
DAA: 22-10-2009 10:00

✓ **baba** 28 *Mx Ni CR Cu RD PR Co Ve Ec Pe Bo Ch Ar Ur Am*
✓ **babá** 2 *Mx*
⊖ **baba blanca** 1 *Gu*
✓ **bababuy** 3 *Co Am*
✓ **babacero** 1 *RD*
✓ **babaco** 5 *Ec Am*
✓ **babacuy** 1 *Co*
✓ **babada** 3 *PR Am*
⊖ **baba de buey** 1 *Ni*
⊖ **baba del diablo** 1 *Ar*
⊖ **baba de loco** 1 *Ur*
⊖ **baba de sapo** 1 *Ch*
✓ **babagüi** 1 *Co*
⊖ **babahoya** 1 *Am*
⊖ **babahoyana** 1 *Am*
⊖ **babahoyano** 1 *Am*
⊖ **babahoyense** 1 *Am*
⊖ **babahoyo** 1 *Am*
⊖ **Babahoyo** 1 *Am*
⊖ **babaidor** 1 *Am*
✓ **babalao** 5 *Pa Cu RD PR Ve Am*
⊖ **babalón** 1 *Ve*
✓ **babalú** 1 *Cu*
⊖ **Babana** 1 *Ve*
⊖ **babanca** 2 *Am*
✓ **babandí** 2 *Ve*
⊖ **babas** 1 *Ch*
✓ **babasfrías** 4 *Mx Co Am*
⊖ **babastibias** 3 *Am*
⊖ **babasú** 1 *Am*
✓ **babay** 4 *RD Co Ve Ec*
✓ **babaza** 1 *Ve*
✓ **babazal** 1 *Ni CR*
⊖ **babazas** 1 *Am*
✓ **babazón** 3 *Ni Co Am*
⊖ **babazorra** 2 *Am*
⊖ **babazorro** 2 *Am*

**baba**: 30 entrada(s)

| **Gen. 10/9** | | **Andes 9/8** | | | | | **Caribe 5/5** | | | | **Central 2/2** | | **Chile 1/1** | **Mexic. 2/2** | | | | | **R. Plata 1/1** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Am. | Esp. | Bol. | Col. | Ec. | Perú | Ven. | Cuba | Pan. | P. Rico | R. Dom. | C. Rica | Nic. | Chile | El Salv. | Guat. | Hond. | Méx. | EE.UU. | Arg. | Par. | Ur. |
| **10**/9 | 0/0 | **1**/1 | **4**/3 | **1**/1 | **1**/1 | **2**/2 | **3**/3 | 0/0 | 0/0 | **2**/2 | **1**/1 | **1**/1 | **1**/1 | 0/0 | 0/0 | 0/0 | **2**/2 | 0/0 | **1**/1 | 0/0 | 0/0 |

Entradas completas de **baba**

**Neves, A. N. (1975) [hi - Am ] *Diccionario de americanismos*. Sopena**

[ **BABA.**]

**BABA.** f. *Venez.* Especie de *caimán,* algo más pequeño que el común y de hocico ancho *(Alli-gator punctatus).* No ataca a las personas.

**Arias de la Cruz, M. A. (1980) [hi - Am ] *Diccionario temático. Americanismos*. Everest**

[ **BABA**]

BABA

*n. f. Zool.* (Colomb., Venez.). Yacaré, especie de caimán o cocodrilo.

**- (1982) [hi - Am ] *Americanismos. Diccionario ilustrado Sopena*. Ramón Sopena**

[ **BABA.**]

**BABA.** f. ***mí*** Especie de caimán, "2 Mariposa grande y de colores brillantes. ■3 Palabrería, dicho insubstancial.

[ **BABA.**]

**BABA.** adj. ***mí*** *Etnog.* pícese del individuo de una tribu de indios del Ecuador que vivía a orillas del río Babahoyo, en territorio que hoy pertenece a la actual ■provincia de Los Ríos. Ü. t. c. s. «2 Perteneciente o relativo a estos indios.

**Richard, R. (coord.) (1997) [hi - Am ] *Diccionario de hispanoamericanismos no recogidos por la Real Academia. (Formas homónimas, polisémicas y otras derivaciones morfosemánticas)*. Cátedra**

[ **baba (o: babilla)**]

**baba (o: babilla).** f. **(1)** Pequeño cocodrilo inofensivo *[Caimán cocodrilus, Jacarépunctulatá)* de ancho hocico, que vive en los ríos y en las lagunas de tierra caliente; la carne de su cola es sabrosa. (Ven. = Col.): «(...) Misia Baba respondió sin pensarlo dos veces: / -Hagan una junta de beneficencia. / Pronunció la frase anterior con un estilo contundente y conciso de sibila, abriendo apenas un instante el ojo para sumirse de nuevo inmediatamente en las profundidades de su beata digestión (...).» (A. Arraiz, *Tío Tigre y Tío Conejo,* 29) = santamaría dga = tejera = consultas **(2) baba del diablo.** f. Filamentos de origen vegetal. (Ur.): «En los cabellos de las chinas* las semillas de sorgo o las babas del diablo hablaban a las claras del idilio gozado...» (E. Amorim, *La carreta,* 82) = consultas

**Morínigo, Marcos A.; Morínigo Vázquez-Prego, M. A. (1998) [hi - Am ] *Nuevo diccionario de americanismos e indigenismos*. Claridad**

[ **BABA**]

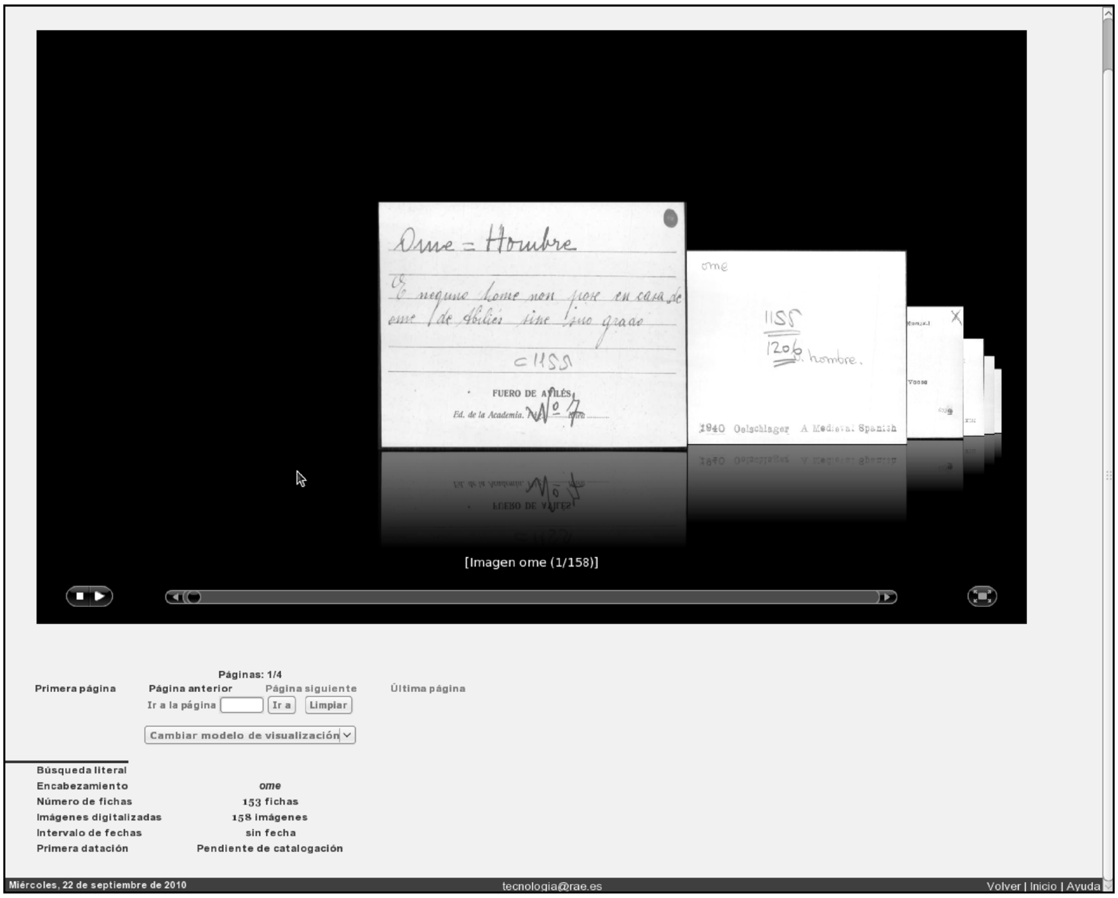
Figure 3: Unified query of American lexical repertoires

Finally, ARU contains about 40,000 digital images (from which the aforementioned textual conversion has been made) and is composed of over 550,000 items.

## 3.2 Lexical and lexicographic applications

The Royal Academy has begun digitalizing its lexical files, starting with the so-called *General file* (*Fichero general*). This file consists of just over 10 million records on index cards that directly reflect the uses of a word – in its various diachronic and dialectal variants – or they have attached an article taken from a dictionary. The oldest index cards contain the examples, or ‘authorities’, that were used in the first dictionary published by the Royal Academy between 1726 and 1739.

Due to the technical difficulty in obtaining reasonable results in optical character reading – since many of these index cards were handwritten or have low contrast or archaic fonts –, the index cards are stored as images, linked to the heading contained therein. For now, we have developed a query interface on this material, with different viewing modes. The following images provide some examples:



Ome = Hombre

E negune home non pose en casa de ome / de Abiliés sine suo grado

c1155

FUERO DE AVILÉS

Ed. de la Academia. Nº 7

ome

1155
1206 v. hombre.

1940 Oelschlager A Medieval Spanish

[Imagen ome (1/158)]

Páginas: 1/4

Primera página Página anterior Página siguiente Última página

Ir a la página [ ] Ir a Limpiar

Cambiar modelo de visualización

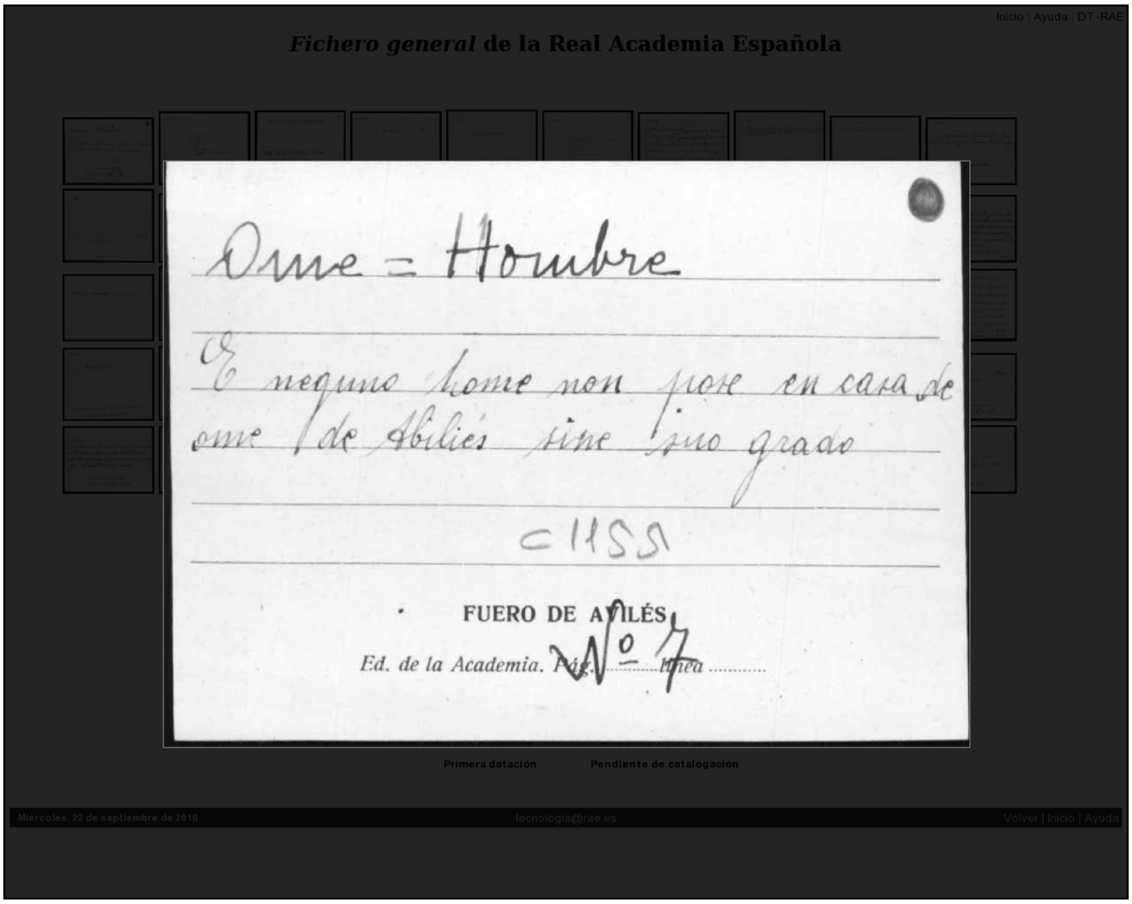
Búsqueda literal

| | |
|---|---|
| Encabezamiento | *ome* |
| Número de fichas | 153 fichas |
| Imágenes digitalizadas | 158 imágenes |
| Intervalo de fechas | sin fecha |
| Primera datación | Pendiente de catalogación |

Miércoles, 22 de septiembre de 2010 tecnologia@rae.es Volver | Inicio | Ayuda

Figure 4: Pictures of index cards of *ome* ‘man’

Figure 4 shows some examples of *ome* (one of the historical variations of the word *hombre*, ‘man’) as a slideshow. Both the horizontal scroll bar and the mouse wheel can be used to see the images, which are organized in batches, or lots, of 50. They can also be displayed in this format as a slide carousel, where users can stop at any point. On the other hand, the following image (figure 5) presents the same information (the first index card) as a mosaic, barely noticeable because it has been maximized. This card, which has a red circle mark, is the first dating of that word found by this institution: as can be seen, *ome* dates from around 1155, and is documented in the *Fuero de Avilés*.



***Fichero general* de la Real Academia Española**

Ome = Hombre

E neguno home non pose en casa de ome / de Abilies sine suo grado

c1155

FUERO DE AVILÉS

Ed. de la Academia. Pág. Nº 7 línea ...........

Figure 5

### 3.3 Lexical applications

Fifteen years ago the Royal Academy began the construction of two corpora that would include the synchronic and diachronic linguistic usage: CREA (*Corpus de Referencia del Español Actual*, 'Present Day Spanish Reference Corpus', with some 170 million words covering the period between 1975 and 1999) and CORDE (*Corpus Diacrónico del Español*, 'Spanish Diachronic Corpus', with almost 300 million covering from the origins of Spanish until 1974). We have to add to these two corpora the CORPES (*Corpus del Español del Siglo XXI*, '21st Century Spanish Corpus', currently with 70 million words but expected to reach 300) and a series of satellite corpora: one organised in terms of specific types of language (CCT, *Corpus Científico-Técnico*, 'Scientific-Technical Corpus'), another in terms of the linguistic level of the target language of the texts that comprise it (*Corpus Escolar*, 'School Corpus', consisting of textbooks from different levels of secondary education), and another in terms of the purpose of the corpus itself (*Corpus del Diccionario Histórico*, 'Historical Dictionary Corpus', CDH). All these corpora, but especially the first three and the last one, are integrated into the so-called Spanish Data Bank.

During the construction of the latter, the usual design criteria for developing these resources have been taken into account, usually at the national level, but without detriment to the transnational dimension of the Spanish language (*Pan-Hispanic*, as it is called in this case). Thus, the proportions awarded to American texts change over time, which indicates the recognition of the growing importance of American Spanish. Consequently, the initial proportions of CORDE, which give a rate of 70% to European Spanish texts (compared to a 30% for American Spanish), become equivalent in CREA, which provides 50% of each variety in this dichotomy; and that rate 70%/30% returns, though in reverse, in the distribution of CORPES (70% of texts from America, 30% from Spain).

The coding of the texts that form each corpus follows the recommendations of the *Text Encoding Initiative* (TEI), both with respect to the encoding format (XML), and in relation to the bibliographic and textual elements to be registered.

Finally, the corpora have been part of speech (POS)-tagged and each word has been assigned a lemma that connects it to an article in the dictionary (when the voice is included in it). The lematization is, in fact, a consequence of disambiguation.

Once analyzed, the texts are indexed to facilitate quick consultation. As in the rest of the applications described in this paper, all the indexing software was also developed by the Spanish Royal Academy. This indexer competes favourably with other corpus indexers, both in efficiency and versatility, as well as in the size of the corpus that can be managed. The query interface on the Spanish Data Bank allows observation of the absolute and relative frequency and the use and dispersion that the search terms offer. It is possible to consult the lemma and the lexical category (along with its morphological features) on the textual form, even combining several of these criteria, including distance bounded operators on a radio that may be directional or not; this way users can run queries on textual elements which are not necessarily contiguous. The application shown here is not the one that will ultimately be integrated into the new website of the Academy, as this is still under construction.

## 4. Conclusion

This document offers an illustrated summary of some of the technological developments undertaken in recent years by the Spanish Royal Academy to improve working conditions for the teams of collaborators attending academics in their language tasks (which may or may not be lexicographic). Again as part of its mission to serve the Hispanic speaking community, the Royal Academy is making an effort to make available these tools that allow speakers to have a better knowledge of their language. Due to the limitations of such a summary, this paper leaves out some developments already underway, such as the Neologism Observatory and, of course, does not dwell on the great technological effort (both in software engineering and language technology) on which applications presented here are supported. The immediate future will allow any user to view these resources through any browser in the world, which in turn will be supplemented with new interfaces and a more profound treatment of those already available.[3]

[3] We would like to thank Dámaso Izquierdo Alegría for his thorough revision and editing of this text.