Dialectologia. Special issue, IV *(2013), 5-24.*
ISSN: 2013-2247

Received 15 July 2013.
Accepted 30 August 2013.

THE *TESOURO DO LÉXICO PATRIMONIAL GALEGO E PORTUGUÉS*. A GALICIAN AND PORTUGUESE WORD BANK[1]

[1]Xosé Afonso ÁLVAREZ PÉREZ & [2]Xulio SOUSA

[1]Centro de Linguística da Universidade de Lisboa / [1, 2]Instituto da Lingua Galega, USC

xoseafonso.alvarez@gmail.com / xulio.sousa@usc.es

Abstract

The international project *Tesouro do léxico patrimonial galego e portugués* (Thesaurus of the Galician and Portuguese heritage lexicon) aims to be a cross-dialectal lexical portal bringing together lexicographical material from Brazil, Galicia and Portugal in a single computer tool. This dialect portal will give direct access via Internet, free of charge, to a large body of lexicographical data much of which has until now remained unpublished and hard for researchers obtain. The lexical information in the *Tesouro* is fully lemmatized, semantically classified and geographically referenced, making it possible to obtain usefully grouped search results and generating a map representation corresponding to each data set. Besides its obvious value to dialect researchers and lexicographers, the *Tesouro* will also provide useful material for the study of names, historical linguistics, etymology, morphology and so on. This tool might also be exploited in ethnographical and historical research since it makes available to the scientific community a large amount of information about the inherited traditions, both material and immaterial, of all three countries, much of which is endangered on account of recent changes in traditional ways of life.

Keywords

lexicography, Galician, Portuguese, lexical dialectology, linguistic geography

[1] This paper falls within the projects *Tesouro Dialectal Português* (Fundação para a Ciência e a Tecnologia, Portugal, PTDC/CLE-LIN/102650/2008) and *Tesoro del léxico patrimonial gallego y portugués. Banco de datos electrónico (corpus gallego) y cartografía automática* (Ministerio de Ciencia e Innovación, España, FFI2009-12110). X. A. Álvarez Pérez received a grant from the human resources program of the Portuguese *Fundação para a Ciência e a Tecnologia*. The authors of this paper have been members of the project since it began.

O *TESOURO DO LÉXICO PATRIMONIAL GALEGO E PORTUGUÉS*.
UN TESOURO LÉXICO DO GALEGO E DO PORTUGUÉS

Resumo

O proxecto internacional *Tesouro do léxico patrimonial galego e portugués* ten como obxectivo constituír un portal de léxico dialectal que reúna nunha mesma ferramenta informática materiais lexicográficos procedentes de Brasil, Galicia e Portugal. Este portal dialectal permitirá o acceso inmediato, através da internet e de modo gratuíto, a unha gran cantidade de datos lexicográficos, moitos deles aínda inéditos e de difícil acceso para os investigadores. A información léxica introducida no *Tesouro* está debidamente lematizada, clasificada semanticamente e xeo-referenciada, polo que é posible obter resultados debidamente agrupados e xerar a representación cartográfica correspondente. Ademais do obvio interese para as investigacións de tipo dialectal e lexicográfico, o *Tesouro* fornecerá material de utilidade para pescudas onomasiolóxicas, de lingüística histórica, etimolóxicas e morfolóxicas, por citar algúns exemplos. Do mesmo modo, esta ferramenta tamén poderá ser aproveitada para estudos de tipo etnográfico e histórico, xa que pon ó alcance da comunidade científica moita información sobre o patrimonio material e inmaterial tradicional dos tres países, moi ameazado polo cambio nas formas de vida ocorridos nas últimas décadas.

Palabras clave

lexicografía, galego, portugués, dialectoloxía léxica, xeografía lingüística

1. The objectives of the *Tesouro*

Tesouro do léxico patrimonial galego e portugués [Thesaurus of the Galician and Portuguese heritage lexicon] (henceforth *Tesouro*) is a joint initiative of Brazilian, Galician and Portuguese universities under the direction of Rosario Álvarez of the University of Santiago de Compostela's Instituto da Lingua Galega. The initial intention of this project was to create a computerized corpus bringing together lexis linked to traditional culture, particularly in connection with activities and areas of knowledge that have been lost or are in the process of extinction owing to cultural and social change. It was also planned that the geographical origin of all data should be specified in order to facilitate comparative studies and research on lexical diffusion. The original idea was to develop a project for Galician materials only, but it was soon realised that

it would be advantageous to broaden the project to include Portugal and Brazil, which are linguistically linked to Galicia.

The product resulting from this project will constitute a website accessible to the whole academic community that provides access to materials now scattered about in different types of sources. A substantial part of the dialect research was carried out in Galician, Portuguese and Brazilian regions through fieldwork performed by university students and usually submitted as undergraduate or doctorate dissertations which were in most cases unpublished. Despite the interest and value of this material, its use in studies on dialect variation or lexical research in general had been limited because the information was difficult to get hold of. Linguistic atlases constitute another type of source whose utilisation will be facilitated by the *Tesouro*. Despite the wealth of data in these works, for many years the way the information was traditionally presented made it difficult to consult. The incorporation of material from linguistic atlases into the project will also allow the data they contain to be recovered with greater ease and thoroughness. Moreover, because of their scope or the manner in which they were published, some of these works on linguistic geography have had a very limited distribution and are therefore resources to which access is difficult for many researchers.

Another purpose of the project is to contribute to the study of the inherited traditions, both material and non-material, of the countries in question, thereby making a significant contribution to ethnographic research. The culture and technology associated with the rural world are extinct in most parts of our countries owing to fargoing economic and social changes in recent decades. Consequently, the lexis associated with traditional activities and culture is also dying out.

The sources incorporated into the *Tesouro* provide a very large body of information about traditional ways of life from various perspectives. First of all there is vocabulary. The papers and publications covered by the project focus on semantic fields linked to traditional culture and in consequence they give very thorough coverage to large areas of the lexicon concerned with different forms of work that used to be performed and also the objects associated with such activities (e.g. grinding flour and making bread, growing flax, ploughs and carts, etc.). Integrating these

sources into a unified corpus makes it possible to fill out the information in the different sources and will provide a basis for building thematic vocabularies on different subjects. Even the definitions for each word given in the glossaries often contain interesting information about material aspects, such as descriptions of tools, their parts, how they are used, related customs and so on. What is more, many sources contain drawings or photographs taken at survey locations; this invaluable material consisting of hundreds of images of many kinds has been digitalized and can be viewed in the *Tesoro* portal together with their respective entries.

The third purpose of the *Tesouro* is to provide a broad lexical corpus that can serve as a basis for different kinds of synchronic and diachronic research. Apart from the obvious interest for dialectology and lexicology, the application will also provide material relevant to various initiatives in other fields, especially etymology, phonetics and phonology,[2] lexicography[3] and morphology. Of particular interest in this respect are contrastive studies between Galician and Portuguese or between European and Brazilian Portuguese, which may through considerable light on processes of linguistic expansion over time, lexical stratigraphy and the circulation of words between different areas. Different processes of diachronic dialectology and so on. The existence of a solid corpus of Galician and Portuguese dialect material will also allow comparisons with the rest of Romance in order to study points of convergence and divergence, particularly regarding the relationship with Romance languages spoken in the Iberian Peninsula.

Yet another of the main aims of the *Tesouro* is to propose a model of a computer tool for setting up a dialectal lexical corpus. For this purpose it will be necessary to address a number of challenges in order to obtain a user-friendly tool. The first challenge is to incorporate in a single data base materials of a variety of types and internal structures which furthermore belong to three different linguistic subsystems with distinct spelling rules. The second challenge has to do with developing an

[2] The *Tesouro* is a project focusing on lexicon; however, its database includes all the phonetic information given in the dissertations, whether this be exhaustive transcriptions in phonetic symbols of every form occurring, or adapted conventional spellings to reflect different phonetic processes that develop in the language.

[3] A study about the incorporation of dialect forms in dictionaries based on the Portuguese material in the *Tesouro* may be seen in Álvarez Pérez (2012).

automated cartography system to display the results obtained by the user as maps on the screen in order to facilitate the exploration of their geographical distribution over a vast area (the combined land area of Brazil, Galicia and Portugal total more than eight and a half million square kilometres). The members of the project aim to produce a computer tool and web portal for the *Tesouro* which will also be able to be used in other linguistic domains (Montemagni & Picchi 1998, Barbato & Varvaro 2004, Kemps-Snijders & Wittenburg 2006, De Vriend *et alii* 2006, Van Keymeulen & De Tier 2010).

Lastly, one of the objectives of the *Tesouro* is to give rise to a dialectal lexical portal which, besides providing an electronic edition of traditional vocabulary, contains complete and varied information of interest to researchers in a wide range of disciplines. Thus in addition to the concept of an open-ended data base which includes photographs, drawings and many types of ethnographic information, the application will also contain an exhaustive inventory of Galician and Portuguese dialect sources including bibliographical information on all items about aspects of lexical dialectology in these areas.

2. The sources

The *Tesouro do léxico patrimonial galego e portugués* proposes to incorporate any kind of study whatsoever with lexicographhical content that includes geographically localised dialect materials from Brazil, Galicia or Portugal. The project will encompass a wide variety of materials of diverse form, external and internal structure and territorial variation. On the basis of the materials so far incorporated, three fundamental types of source can be distinguished:

a) Ethnolinguistic monographs on the speech variety of a particular town or small area

These were mainly conceived of as supervised academic essays aiming to study the language of a small area, normally a parish, a village or a cluster of villages as in Santos' (1967) cross-border study. They are usually concerned with describing the grammar and lexicon of the area in question or are monographic lexical studies.

Typically they cover several semantic fields, with particular attention to traditional topics, though there are also studies that focus on specific domains such as the language of fishermen (Alves 1958), bakers (Machado 1949) or potters (Vieira 1960).

Most of these studies are difficult for researchers to get hold of as they are usually unpublished and must be obtained either from the authors or in the library of the institution where they were produced or the university where the thesis was submitted. They include an interesting set of studies from the thirties and forties of the last century and such studies have significant conservation issues. The incorporation of these sources in the *Tesouro* not only facilitates the scientific community's access to them without needing to move the originals around but also ensures the survival of their content regardless of the fate of their physical support.

b) Language atlases and dialect surveys

The group just mentioned involves dialect studies performed in a specific place or a narrowly delimited geographical area; another set of studies concerns fieldwork data (obtained directly or by correspondence) surveying a much wider network of localities spread across one or more of the countries covered by the *Tesouro*. Linguistic atlases are of particular importance within this category. The *Tesouro* will incorporate material from published geolinguistic studies but also from previously unpublished data in atlases that are currently being published or developed by groups in which there are members of the *Tesouro* team participating.

Of special interest is the fact that the *Tesouro* corpus will include data from several large-scale atlases covering the whole country (or state, in the case of Brazil). One of these is the *Atlas lingüístico galego* (ALGa), which is being published by the Instituto da Lingua Galega, the *Atlas Linguístico-Etnográfico de Portugal e da Galiza* (ALEPG), which is being developed by the Centre for Linguistics of the Universidade de Lisboa, and the *Atlas lingüístico de la Península Ibérica* (ALPI), only one volume of which was ever published, in 1962, but which is now being prepared for publication in its entirety. The surveys for the *Atlas linguístico do Brasil* are still ongoing, but Brazil is covered by a good number of regional atlases that are being incorporated into the *Tesouro* as well as the *Atlas Linguístico do Estado do Ceará* (ALECE) and the *Atlas*

Lingüístico da Paraíba (ALP). Although not strictly an atlas, we should also mention the material comprising the Universidade de Coimbra's *Inquérito Linguístico Boleo* (ILB), a monumental corpus of more than 3000 surveys mainly obtained by correspondence. Besides these generic atlases, we also have a number of studies focusing on more specific subject areas. One area of particular importance is sailors' vocabulary, for which there are three studies of particular significance: the *Atlas linguístico do litoral português* (ALLP) by Gabriela Vitorino, *Nomenclatura de la fauna y flora marítimas de Galicia* by Carme Rios Panisse (Rios Panisse 1977, 1983) and the sections for Galician and Portuguese localities of the *Léxico de los marineros peninsulares*, directed by Manuel Alvar.

Incorporating the materials in these linguistic atlases into the *Tesouro* will greatly benefit geolinguistic and lexicographical research in several ways. One is by making it possible to consult a wide range of unpublished data (it is a well-known fact that publication of linguistic atlases is a long, drawn-out process). Another is by making it much, much easier to check on data. Traditional linguistic atlases are organised in such a way that users have to pore through all the maps one by one in search of the form that concerns them and the geographical location that interests them, especially in atlases that lack exhaustive indices (Alvar López 1982, Castillo Peña 1990, Le Dû 1997, Montemagni & Picchi 1998, Ronco 2004, Sousa Fernández 2010). The *Tesouro* will give quick and easy access to all these materials.

c) Other sources with dialect lexicon

In the domain of Galician and that of European Portuguese, there exist many studies of aspects of local culture from an onomasiological or ethnolinguistic point of view. They usually take the form of articles in academic journals or monographs. Some took their inspiration from the *Wörter und Sachen* school of philology, which focused on the exhaustive study of the lexicon of material culture in different parts of the Iberian Peninsula during the first half of the twentieth century (Ebeling 1932; Krüger 1963; Schneider 1938). Another example of this type is the prolific work of the Galician scholar Xaquín Lorenzo (Lorenzo 1982a, 1982b, 1982c, 1983), who studies practically

every area of material culture in Galicia in studies published in the first half of the twentieth century.

In addition to these works which are well known and fully accessible to specialists, there are numerous others which are worth rescuing and making better known. These include contributions on subjects of local or regional interest, generally published in local journals with limited circulations. Their quality varies greatly, but they often represent invaluable and practically unique sources for the study of local lexicon. One example of a recent study of this kind that might be mentioned here is Vila Fariña (2005). Of late, many resources are appearing which take advantage of the possibilities of new technology to publish lexicographical collections on the Internet. Some examples illustrating this may be found at http://escavar-em-ruinas.blogs.sapo.pt/tag/dicion%C3%A1rio, with 21 blog entries published between 2008 and 2011, with entries of a *Dicionário de Falares do Minho*, and on the web site at http://historiaselendas.no.sapo.pt/paginas/falar.htm about the speech variety of Cuba in the Alentejo region of Portugal presenting a collection of words and phrases considered typical of the area.

3. Treatment of materials and structure of the database

From the very beginning when designing the project it was decided that the computer application providing access to *Tesouro do léxico patrimonial galego e portugués* should hold all the information contained in the original sources while at the same time allowing this information to be accessed in a handy and useful manner that responded to users' needs. Merely reproducing the content of each source literally would not be enough; it was essential to treat the data meticulously and organise it in such a manner as to facilitate different kinds of search and sorting or presentation of search results. Despite the challenging nature of this work of organisation and classification of information given the heterogeneity of the materials, the structure of the data base is sufficiently rich and flexible to accommodate all the information in the sources.

There now follows an outline of the main fields making up the *Tesouro* data base's structure which are used to standardise and sort the information:

a) *Variant*. The form that stands at the beginning of an entry in each of the glossaries presented in the *Tesouro*. The spelling variant used by the author is scrupulously respected since it often provides useful information, especially from a phonetic point of view (e.g. Pt. *cereja*, *cereija, ceraija, saraija,* etc. 'cherry'); the heterogeneity of a range of variants is sorted and grouped through attribution of a single headword (lemma), which coincides with the standard form in the language (which in the example just given would be *cereja*), as we shall see in section d). When a form that has been collected is only given in phonetic transcription in the original source, the corresponding variant in conventional spelling has to be created.

b) *Phonetic transcription*. Although the focus of the project is lexicographic, it was thought worth retaining phonetic information whenever any was given in the sources. Since the conventions used to represent the pronunciation of words vary widely, it seemed essential to unify and adapt phonetic transcriptions using API symbols. This adaptation was carried out in a way that aims to conserve the most pertinent information found in the sources.

c) *Part of speech*. The lexical category given in the source is kept *as is*, formulated as in the original. Thus the same word might be categorised in different places as *s, sm, subst, subst m*, etc. As an aid for classifying and sorting materials of different kinds, a standardised part of speech category is also supplied for the assigned lemma, and placed in a special field in the data base. The reason for keeping the information shown in the original work is to provide an opportunity to compare the categories of variants with those of the headwords, which is useful, for example, when studying gender or number change phenomena (cf. *o febre, as tomates*).

d) *Headword (lemma)*. Different phonetic or phonological variants found in sources are brought together under a single headword which makes it easier to see all such variants at a glance. The headword is distinct for each of the language varieties (Galician, European Portuguese and Brazilian Portuguese), so for example the variant forms *dereito*, *dreito*, *direito* in Galician are grouped under the headword DEREITO and in Portuguese under DIREITO. Forms considered to be made up of derivative morphemes

are treated as separate headwords (e.g. QUEIXO, QUEIXELO, QUEIXAL). In order for users to be able to access the variants compiled from sources in all three linguistic areas starting from specific headwords, links were established between the three lemma banks. Therefore when one looks up the Galician headword DEREITO it will be stated in the search result table that the corresponding Portuguese headword is DIREITO and the user will have the option of looking up all the variants assigned to both headwords.

e) *Examples*. In some cases sources provide examples of the use of forms. Sometimes grammatical information such as government is provided, other times collocational information is given, and very often verse fragments or idiomatic expressions that contain the word under consideration are cited. It is also fairly common for the example given to represent a sample of the living speech of the locality, which may be useful for purposes beyond strictly lexical study.

f) *Cross-references*. It is quite usual in some of the sources for an entry to include cross-references to related items occurring in the same source. Such cross-references may serve, for instance, to link two formally related variants of the same word (e.g. *albitanas/albitanes*); or the cross-reference may indicate a meaning relationship, for example of hyponymy or heteronymy (e.g. *jugo ~ tchabielha, canzile, molida, solada, solinho, temoeiro, canga...*); yet again, cross-references may link a number of forms considered secondary to a primary entry which contains in one place all the information about meanings, examples and other aspects. In all these cases, it was considered necessary to conserve this network of links, which our application allows users to follow up with ease without needing to start a new search.

g) *Definitions*. Into this field goes semantic information, considered the fundamental part of the sources we compile. *Definition* is to be understood in a broad sense, including not only the components normally thought of as a definition in the narrow, dictionary sense, but also information that may be interpreted as explaining the meaning of the item in question. It is not unusual, in some monographs about local Galician speech varieties, to find a section on the lexicon where forms are not indexed but rather occur within a descriptive text that discusses a particular semantic field and the principal characteristics of the elements that form part of this. Information that does not fit into other, more specific fields are also put in this field, such as footnotes

(which are nearly always bibliographical references), geographical indications about the place where the variant was collected or observations about whether or not the form is listed in the dictionaries (frequent in the Brazilian and Portuguese materials).

h) *Semantic classification*. Several of the sources incorporated into the *Tesouro* organise lexical information by semantic fields. This is not done in all sources and rarely is it based on common criteria or shared theoretical principles. When designing the present project it was found desirable to carry out a data-oriented semantic classification of all the incorporated materials. This classification should make it possible to extract information grouped by semantic fields, such that users can obtain a listing of all words linked to a single semantic cluster (such as weather, types of agricultural terrain, plants and trees, buildings, etc.). A system of semantic classification was developed to this end based on earlier studies which results in twelve major types, which may be revised in the future to create subdivisions.

i) *Geographical index*. One of the conditions that must be met by materials to be incorporated into the *Tesouro*'s data base is geographical specification. This requirement implies the attribution of all lexicographical data to a village, parish, municipality or other identifiable geographical entity. Between the two options for indicating the origin of forms, geographical point or area, the latter was chosen because it admits of mapping and gives an idea at a glance of the distribution of forms across the three countries' territories. In view of the differences in size between the three territories, we agreed to use different administrative entities to represent the data. For Galicia and Portugal the *concello* or municipality is used as the entity of reference. For Brazil we chose the *mesorregião*, an administrative division covering several municipalities with similar economic and social characteristics.[4] Every variant in the data base is referenced by a code to the administrative entity to which the place where the item was collected belongs. If more specific information is given in the source document about the particular place where the item was collected, this is also indicated in the text file which users may access, and in many cases this specification also appears in the "Definition" field.

[4] A glance at the map of *mesorregiões* shows that their size is appropriate given the population density in Brazil.

j) *Pictures and drawings*. Many of the materials that have gone into the *Tesouro* have graphic content serving to illustrate the objects that are described and defined. These photographs and sketches are also placed in the data base and can be consulted at the same time as the textual information is being accessed. These illustrations were digitalized and their quality was even improved so that users can fully benefit from the additional information they provide.

4. The query tool

The *Tesouro*'s query application is available on a publicly accessible web page free of charge. The task of designing this page and the query-processing application on it has turned out to be one of the most laborious and time-consuming parts of the whole project. Although many existing tools of a similar kind served as a guide and model, the characteristics of the original sources, the fact that we are dealing with three linguistic domains and the need to handle both textual and graphic information at the same time led to longer delays in the development of the application's design than had originally been anticipated.

The main purpose of our project is to give researchers access to data sources with lexicographical data that are unpublished or difficult to consult. Data collected by dialectologists in different areas from different informants may be classified and organised in different ways, with dictionaries and dialect databases representing a crucial part.

In order to make lexicographic data exploitable from different perspectives we organised original information into two main types: lexicographical information and geolinguistic information. All the sources provide information about meaning and information about where the word is used (geographical information). The first step in our project is to organise the information in a database structure where each language item (word or idiom) is characterised with respect to these two dimensions. In the *Tesouro* the lexicographical information is shown in text format and the geolinguistic information is displayed using text format and a cartographic representation, with

areal thematic maps. The application admits searches starting from a *headword* or a *variant*. A headword search produces a table containing all the variants of the headword with information about each variant. The result of a variant search is a list of all variants identical to the search item with information about these.

The search results are displayed in two blocks: lexicographical information (text) and geolinguistic information (maps).

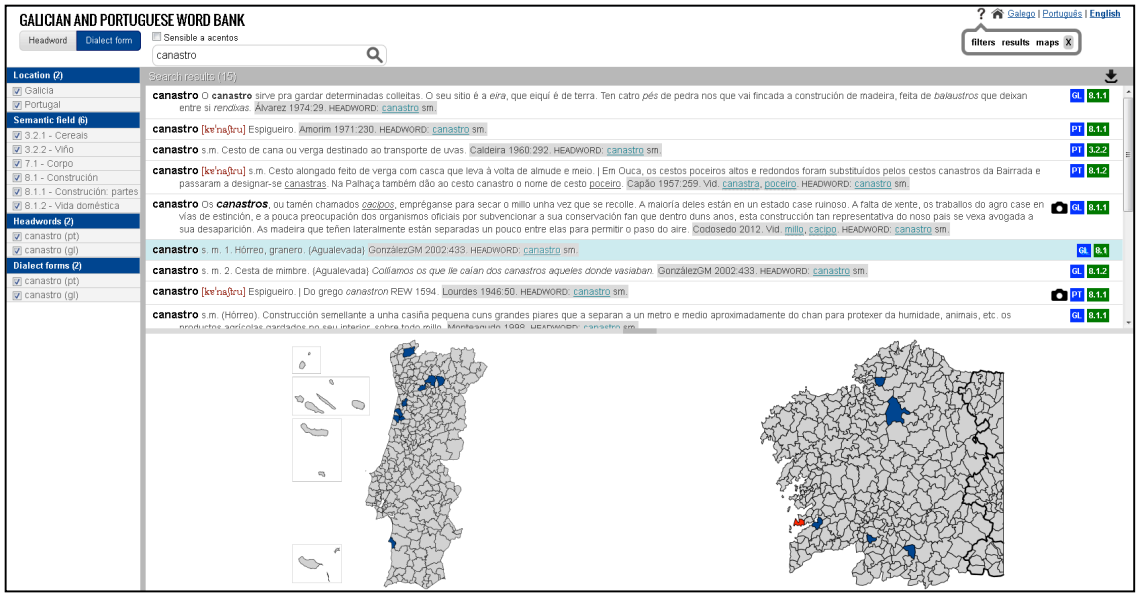

Figure 1. Simple query: *canastro* (dialect forms)

4.1 Lexicographical information

Textual information taken from the sources is displayed in a frame at the top right below the search box. The heading of the frame indicates the number of results found for the search item (21 entries for the headword CANASTRO in Figure 1; 15 entries for the variant *canastro* in Figure 2). The textual information for the search appears within the frame organised as follows. Each horizontal row represents a record in the data base. In the heading of each row the variant is set off in bold, in the spelling found in the source, and is followed by the corresponding lexicographical data: phonetic transcription, part of speech, meaning, examples and other information. Following this information, all of which comes from the source, further information added in the course of data processing is shown against a grey background: the reference for the

source, a listing of other items in the data base and in the original work, the headword to which this variant has been assigned, and the part of speech of the headword. At the far right of the row three coloured icons indicate whether there is a picture linked to the variant, the area where the data were collected and a code specifying the semantic class assigned to the variant. Clicking on the camera icon displays the linked picture (see Figure 3 for the variant *canastro* in the fifth row). All the information contained in this frame can be downloaded as a text file by clicking on the arrow on the right.

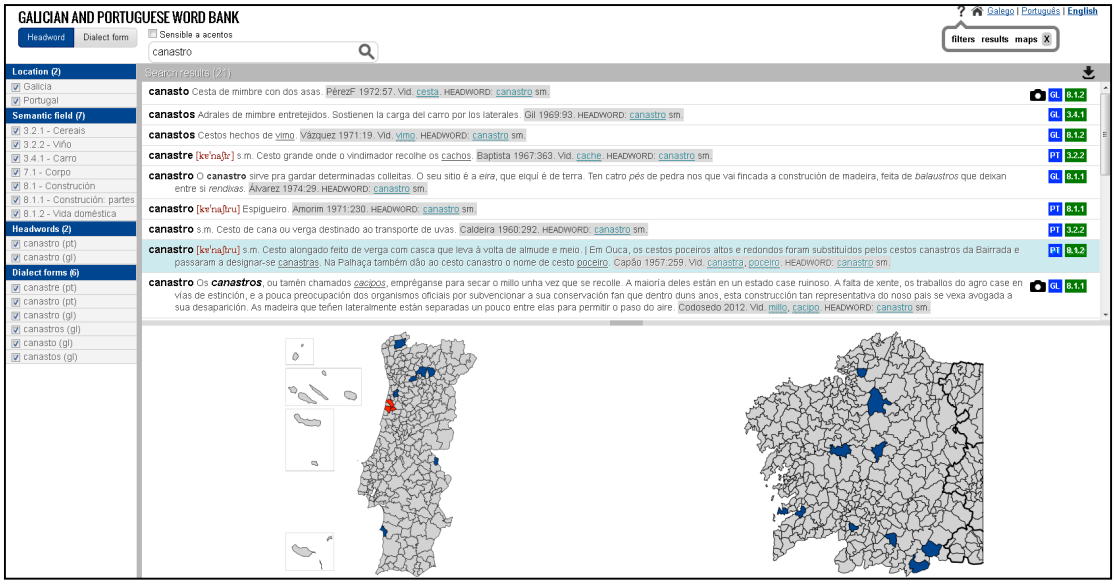

Figure 2. Simple query: *canastro* (headword)

Figure 3. Image: *canastro*

Dialectologia. Special issue, IV *(2013), 5-24.*
ISSN: 2013-2247

4.2 Geolinguistic information

The second major type of information that the *Tesouro* website provides is geolinguistic information. This information is mainly provided in the form of maps in the page's lower frame through an areal thematic map. When a general data base query has been made, this frame displays each of the territories in which variants in the search results box are recorded. In the case of the results for the headword CANASTRO, it turns out that variants of this headword were recorded in Galicia and Portugal. The specific municipalities in each territory concerned are shown in blue. To find out which place corresponds to each variant given in the table all we need do is hover the mouse over a row, and this causes the corresponding municipalities on the map to be illuminated. When the mouse hovers over the area of a municipality on the map, its name appears in a tooltip. The map frame may be expanded or hidden by placing the mouse between the two frames and dragging the separator bar up or down.

Figure 4. Map: Distribution of *canastro* in Portugal

4.3 Filters

In addition to these two frames which display search results (textual information above, maps below), the panel on the left of the page contains a summary of the search results, showing in which countries the item was found, which semantic fields it is associated with and the headwords and variants to which it links (indicating the language to which each form belongs in parentheses). These classifiers may also be used as filters to select, by broadening or narrowing the search, what information is to be displayed in the frames on the right. Filters can apply to the four categories of information mentioned: geographical area, semantic field, headwords and variants. When we click on any of these, the corresponding information on the right toggles between being displayed or hidden. If for example we select only the semantic field *7.1. Corpo* [Body], we shall see that there exists a form *canastro* that is used in the meaning of "O tronco humano" [Trunk of the human body] recorded in the Portuguese municipality of Oliveira de Azeméis (Figure 4). In the heading of each filter the number of categories identified for the current search is shown in parentheses.

All the information in the data base can be accessed directly through an advanced query and filtered by the criteria just mentioned; thus for example variants and headwords may be ordered alphabetically or all the data for a given territory can be displayed (Figure 5).

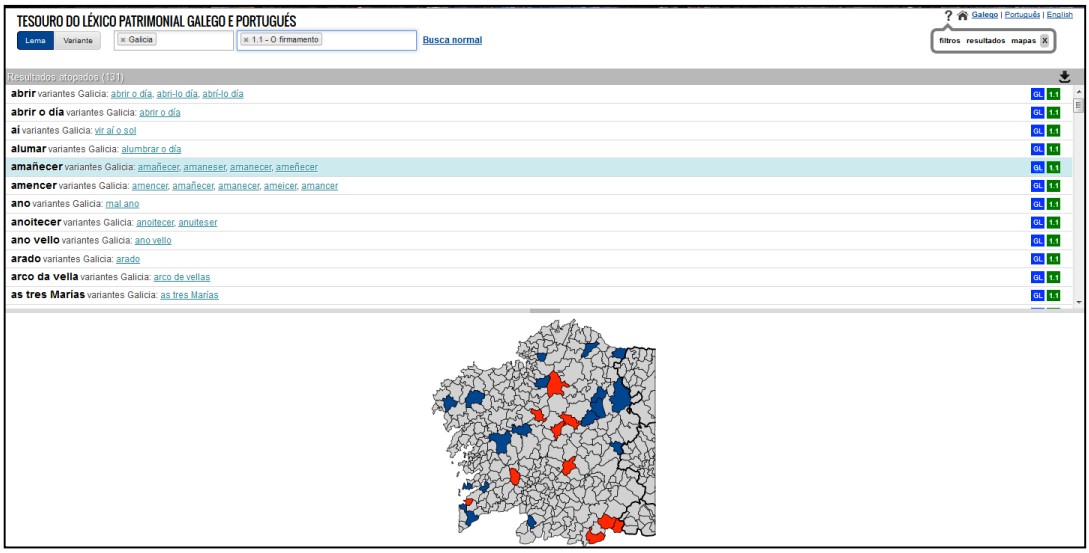

Figure 5. Advanced query: *the sky (Galicia)*

The top right corner of the application's screen gives access to contextual help pages for each of the tools components. The tool also allows us to get complete information about the source materials in the data base and view a fairly exhaustive list of studies of the lexicon of Galician, Portuguese and Brazilian dialects.

5. To conclude

We began by saying that the *Tesouro* project's fundamental objective is to offer a tool for accessing the lexicographical and geolinguistic information found in a large, heterogeneous range of lexicographical works that are difficult to get hold of yet of great interest for various domains of linguistic study, and even other disciplines such as ethnography. Despite some early obstacles for the project, the tool is now at a stage of development such that it is fair to say that few aspects of its structure will need to be modified in the future in order to accommodate new materials later. From the last quarter of 2013, it will become possible for members of the public to use the application to perform queries, particularly on Galician and Portuguese materials, since these were the first languages to enter the project (at present there are 48 sources from Portugal and 35 for Galicia).

We trust that the model we have designed in the *Tesouro* project will serve as a forerunner for similar initiatives in other linguistic domains in the future. We also hope that the *Tesouro* will serve to demonstrate convincingly the great wealth of materials to be found in almost a century of dialect studies in our countries.

References

AHUMADA LARA, Ignacio (2000) "Nuevos horizontes de la lexicografía regional", in Stefan Ruhstaller & Josefina Prado Aragonés (eds.), *Tendencias en la investigación lexicográfica del español,* Huelva: Universidad de Huelva, 15-35.

ALECE = BESSA, José Rogério Fontenele (2010) *Atlas Linguístico do Estado do Ceará*, Fortaleza: UFC.

ALGa = GARCÍA, Constantino & Antón SANTAMARINA (eds.) (1990-) *Atlas Lingüístico Galego*, A Coruña: Fundación Pedro Barrié de la Maza.

ALIAGA JIMÉNEZ, José Luis (1999) "Diatopic labelling in Spanish lexicography: A critical revision and new perspectives", *Cahiers de lexicologie*, 75, 129-152.

ALLP = VITORINO, Gabriela (1987) *Atlas linguístico do litoral português: fauna e flora.* Dissertation in Portuguese Linguistics for the category of Research Assistant. Unpublished, 2 vols.

ALP = ARAGÃO, Maria do Socorro SILVA de (1985) *Atlas Lingüístico da Paraíba: Cartas léxicas e fonéticas*, Brasília: UFPB.

ALPI = CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS (1962) *Atlas Lingüístico de la Península Ibérica. Vol. I: Fonética, 1,* Madrid, Consejo Superior de Investigaciones Científicas.

ALVAR LÓPEZ, Manuel (1982) "Atlas lingüísticos y diccionarios", *Lingüística Española Actual*, IV-2, 253-324.

ÁLVAREZ, Rosario, Xosé Afonso ÁLVAREZ PÉREZ, João SARAMAGO & Xulio SOUSA (2009) "Presentación de un corpus digital de léxico tradicional: Tesouro do léxico patrimonial galego e portugués", *Fonetică şi Dialectologie*, 28, 5-19.

ÁLVAREZ PÉREZ, Xosé Afonso (2012) "O léxico do namoro no Tesouro Dialectal Português (TEDIPOR)", in *XXVII Encontro Nacional da APL. Textos selecionados*, 540-554.

ALVES, Joana Luiza Matos Ribeiro Lopes (1958) *Linguagem dos pescadores da Ericeira*. Unpublished undergraduate dissertation submitted to the Faculty of Letters, Universidade de Lisboa.

BARBATO, Marcello & Alberto VARVARO (2004) "Dialect dictionaries", *International Journal of Lexicography*, 17, 4, 429-439.

CASTILLO PEÑA, Carmen (1990) "Del atlas lingüístico al diccionario: experiencias lexicográficas", in *Actas del Congreso de la Sociedad Española de Lingüística. XX Aniversario* (Tenerife, 2-6 April, 1990), Madrid: Editorial Gredos, 1, 363-371.

DE VRIEND, F., L. BOVES, H. VAN DEN HEUVEL, R. VAN HOUT, J. KRUIJSEN & J. SWANENBERG (2006) "A Unified Structure for Dutch Dialect Dictionary Data", in *Proceedings of The Fifth*

International Conference on Language Resources and Evaluation (LREC 2006), Genoa: Italy.

EBELING, Walter (1932) "Die landwirtschaftlichen Geräte im Osten der Provinz Lugo (Spanien). Sach - und wortkundliche Untersuchungen", *Volkstum und Kultur der Romanen Sprache*, V/1-3, 50-151

KEMPS-SNIJDERS, M. & P. WITTENBURG (2006) "LEXUS – a web-based tool for manipulating lexical resources", in *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*, Genoa: Italy.

KRÜGER, Fritz (1963) *El mobiliario popular en los países románicos*, Coimbra, Universidade de Coimbra, Faculdade de Letras, Instituto de Estudos Românicos, (= Supplement, *Revista Portuguesa de Filologia*, III)

LE DÛ, Jean (1997) "La disparition du groupe des atlas et l'avenir de la géographie linguistique", *Le français moderne*, 65-1, 6-12.

LMP = ALVAR LÓPEZ, Manuel (ed.) (1985-1989) *Léxico de los marineros peninsulares*, Madrid: Arco libros, 4 vols.

LORENZO, Xaquín (1982a) *A terra*, Vigo: Galaxia.

LORENZO, Xaquín (1982b) *A casa*, Vigo: Galaxia.

LORENZO, Xaquín (1982c) *O mar e os ríos*, Vigo: Galaxia.

LORENZO, Xaquín (1983) *Os oficios*, Vigo: Galaxia.

MACHADO, Maria de Lourdes Vaz (1949) *O fabrico caseiro do pão em diversas aldeias do Minho. Subsídios para o seu estudo linguístico-etnográfico*, Unpublished undergraduate dissertation submitted to the Faculty of Letters, Universidade de Coimbra.

MONTEMAGNI, Simonetta & Eugenio PICCHI (1998) "From a Computational Linguistic Atlas to Dialectal Lexical Resources", *Proceedings della Conferenza EURALEX'98*, Liège (Bélgica). <http://serverdbt.ilc.cnr.it/altweb/eurfin98.pdf>

PEREIRA, Maria Fernanda Afonso Alves (1970) *O falar do Soajo*, Unpublished undergraduate dissertation submitted to the Faculty of Letters, Universidade de Lisboa.

RIOS PANISSE, María del Carmen (1977) *Nomenclatura de la flora y fauna marítimas de Galicia. Vol. 1, Invertebrados y peces, con anotaciones etimológicas por Antonio Santamarina*, Santiago de Compostela, Universidade de Santiago de Compostela. (Appendix 7 of *Verba. Anuario Galego de Filoloxía*).

RIOS PANISSE, María del Carmen (1983) *Nomenclatura de la flora y fauna marítimas de Galicia. Vol. 2, Mamíferos, aves y algas*, Santiago de Compostela, Universidade de Santiago de Compostela. (Appendix 19 of *Verba. Anuario Galego de Filoloxía*).

RONCO, Giovanni (2004) "Au delà des dictionnaires: les atlas linguistiques", *International Journal of Lexicography*, 17, 4, 441-455.

SANTOS, Maria José de Moura (1967) *Os falares fronteiriços de Trás-os-Montes*. Separata, *Revista Portuguesa de Filologia*, vols. XII, XIII and XIV.

SARAMAGO, João (2006) "O Atlas Linguístico-Etnográfico de Portugal e da Galiza (ALEPG)", *Estudis Romànics*, XXVII, 281-298.

SCHNEIDER, Hans-Karl (1938) "Studien zum Galizischen des Limiabeckens (Orense, Spanien)", *Volkstum und Kultur der Romanen Sprache*, XI/1-2, 69-145; XI/3-4, 193-281.

SOUSA FERNÁNDEZ, Xulio (2010) "Entre el atlas lingüístico y el diccionario. Un diccionario de léxico tradicional a partir de los materiales del ALPI", in Ignacio Ahumada (ed.), *Metalexicografía variacional. Diccionarios de regionalismos y diccionarios de especialidad*, Málaga: Universidad de Málaga, 237-256.

VAN KEYMEULEN, Jacques & Veronique DE TIER (2010) "Pilot project: a dictionary of the Dutch dialects", in Anne Dykstra & Tanneke Schoonheim, *Proceedings of the XIV Euralex International Congress*, 754-763.

VIEIRA, Carolina Lucília da Silva (1960) *A olaria no distrito de Braga. Estudo linguístico-etnográfico,* Unpublished undergraduate dissertation submitted to the Faculty of Letters, Universidade de Coimbra.

VILA FARIÑA, Xosé Lois (2005) *O falar da nosa aldea: Voces e xiros do falar das xentes de Baión*, Baión: Asociación Cultural O Castro, CD-ROM.