



Corpus de Referencia do Galego Actual
(CORGA)

Guía de uso da aplicación de consulta

Versión 4.1 2024

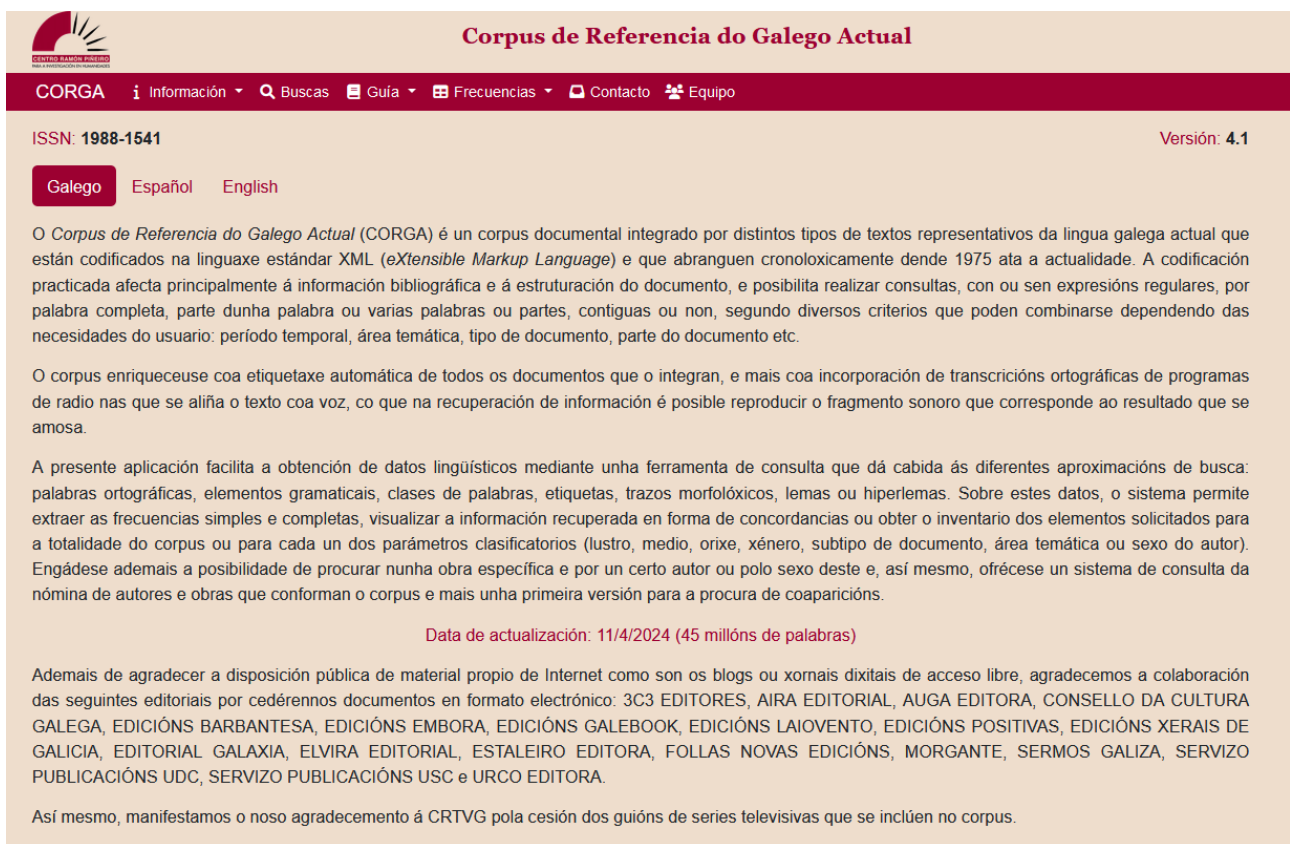
Táboa de contidos

1. Introducción.....	3
2. O corpus.....	7
3. O sistema de buscas.....	9
4. A descarga dos resultados.....	16
5. Buscas simples e buscas avanzadas.....	18
5.1. Metacaracteres e operadores booleanos.....	18
5.2. Sensibilidade.....	19
5.3. Distinción entre palabra ortográfica, elemento gramatical, lema e unidade.....	20
5.4. Consulta por elementos sucesivos.....	27
5.5. Consulta por proximidade.....	28
5.6. O hiperlema.....	30
5.6.1 O hiperlema nos casos de lematización automática.....	36
5.7. Inventario.....	37
5.8. Coaparicións.....	44
5.9. Nómina.....	46
6. Filtros.....	47
6.1. Período cronolóxico.....	47
6.2. Parte estrutural do documento.....	48
6.3. Medio.....	48
6.4. Sección.....	48
6.5. Área temática.....	49
6.6. Autor.....	50
6.7. Sexo do autor.....	51
6.8. Sexo do interlocutor.....	52
6.9. Documento.....	52
6.10. Obra.....	53
6.11. Clasificación textual. O parámetro <i>Tipo de texto</i>	54
6.9.1. Orixe.....	54
6.9.2. Bloque.....	55
6.9.3. Xénero.....	55
6.9.4. Subtipo.....	55
7. Notas para a interpretación dos resultados.....	56

1. Introducción

A presente aplicación dá cabida ás diferentes aproximacións de busca: consulta por palabras ortográficas, elementos gramaticais, clases de palabras, etiquetas, lemas ou hiperlemas, tanto no subcorpus desambiguado manualmente como no *Corpus de Referencia do Galego Actual* etiquetado automaticamente.

A pantalla inicial da aplicación ofrece un texto xenérico no que se presenta o corpus e organiza as funcionalidades e información dispoñible en varias seccións ás que se accede premendo en cada unha das pestanas:



The screenshot shows the home page of the CORGA application. At the top left is the logo of the Centro Ramón Piñeiro. The main title is "Corpus de Referencia do Galego Actual". Below the title is a navigation bar with links for "CORGA", "Información", "Buscas", "Guía", "Frecuencias", "Contacto", and "Equipo". The page includes the ISSN number 1988-1541 and the version number 4.1. There are language selection buttons for "Galego", "Español", and "English". The main content area contains a detailed description of the corpus, its structure, and the search application. It mentions that the corpus is integrated from various text types and is encoded in XML. It also notes that the application facilitates linguistic data retrieval through a search tool. At the bottom, there is a date of update: "Data de actualización: 11/4/2024 (45 millóns de palabras)".

Fig. 1. Pantalla de inicio da aplicación.

Equipo e Contacto son transparentes, polo que nos cinguiremos ás outras catro pestanas: *Información*, *Buscas*, *Guía* e *Frecuencias*, centrándonos logo nas posibilidades que ofrece o sistema de consulta, ao que se accede en *Buscas*.

Na pestana *Información* desprégase un menú que se inicia coa *Descrición xeral* do CORGA etiquetado automaticamente, onde se dá conta dos criterios empregados na selección dos documentos que se incorporan, seguido polos *Datos*, dende onde se accede á listaxe dos documentos que constitúen o corpus e se inclúe a súa distribución segundo os diferentes parámetros de clasificación que se teñen en conta. Deseguido, e en paralelo, ofrécese a información respectiva sobre o subcorpus desambiguado manualmente, explicando a razón da súa existencia e dando conta da súa composición e distribución.

Os ítems *Historial de versións*, *Preguntas frecuentes*, *Ligazóns de interese*, *Documentación*, *Traballos que nos referencian* e *Como citar o corpus?* recollen diversa información de utilidade para o usuario:

Historial de versións sintetiza as distintas versións que existiron do corpus, así como o número de palabras e características xerais de cada unha delas.

En *Preguntas frecuentes* procuramos dar resposta ás cuestións máis usuais en relación coa consulta do corpus, mentres que en *Ligazóns de interese* lístanse corpus, etiquetados ou non, das distintas linguas peninsulares.

Documentación ofrece algunhas publicacións que describen aspectos varios do corpus ou da súa etiquetaxe, en tanto que en *Traballos que nos referencian* se relacionan os traballos que empregaron, e así o explicitan, o CORGA como fonte para o seu estudo.

Finalmente, en *Como citar o corpus?* amósase o modelo de cita aconsellable para referencialo.

Baixo a pestana *Guía* intégranse este manual de uso, no que se describe polo miúdo o funcionamento do sistema de recuperación e extracción de información (*Guía de uso*), unha pequena descrición sobre a codificación e estruturación que sofren os documentos ao integrarse no corpus (*Codificación e estruturación*), a relación das marcas ou etiquetas que se empregan nos textos para codificar fenómenos de distinto tipo (*Etiquetas de codificación*) e, por último, o etiquetario morfosintáctico que subxace na etiquetaxe do corpus, desagregado en *Táboa descritiva* e mais *Exemplos*. Para todo o relacionado cos *hiperlemas* remitimos ao apartado específico que se atopa máis adiante: [5.6. O hiperlema](#).

Na pestana *Frecuencias*, partillada segundo o corpus que escollamos, etiquetado automaticamente ou etiquetado manualmente, poden consultarse ou descargarse:

- as mil formas, elementos gramaticais, lemas e hiperlemas máis frecuentes
- as cinco mil formas, elementos gramaticais, lemas e hiperlemas máis frecuentes
- as listaxes completas das formas, elementos gramaticais, lemas e hiperlemas
- a listaxe de etiquetas
- os datos xerais dos elementos gramaticais e mais os lemas por clase de palabra segundo os parámetros de clasificación medio, xénero, bloque, lustro e área temática.

Por outra banda, entre as novidades que incorpora a versión 4.1 do CORGA, cabe destacar un **Dicionario de frecuencias léxicas** que toma en consideración o índice de dispersión, de maneira que o grao en que os lemas se distribúen nos diferentes tipos de textos, isto é, a súa dispersión, se converte nunha medida moito máis reveladora da relevancia dos elementos léxicos que a súa frecuencia global e normalizada. Para máis información véxase a descrición dispoñible no propio apartado.

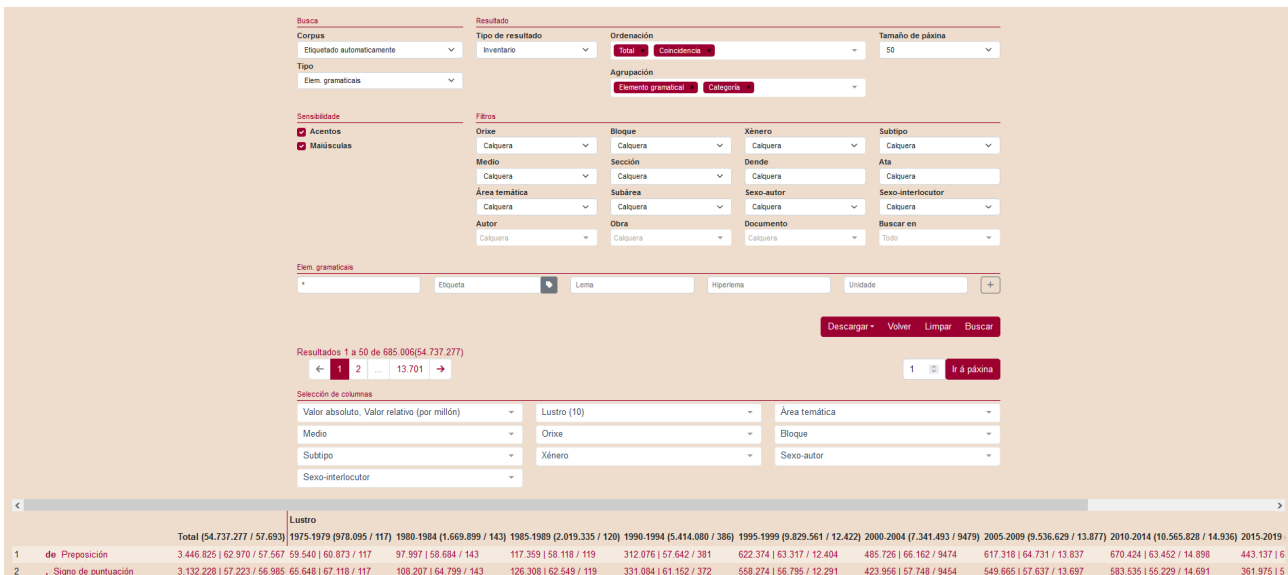
Así mesmo, a través do ítem **Listaxes personalizadas**, situado tamén baixo a pestana *Frecuencias*, proporcionamos un acceso directo e sinxelo ao inventario completo da frecuencia dos elementos gramaticais e dos lemas que conforman o corpus e mais das frecuencias destes segundo os diferentes parámetros de clasificación: lustro, área temática, medio, orixe, bloque, subtipo, xénero e sexo do autor.



The screenshot shows the CORGA website interface. At the top, there is a navigation bar with 'CORGA' and several menu items: 'Información', 'Buscas', 'Guía', 'Frecuencias', 'Contacto', and 'Equipo'. Below the navigation bar, there is a text block explaining the search process. Underneath, the text 'Etiquetado automaticamente' is displayed. The main content area is divided into two columns: 'Elementos gramaticais' and 'Lemas'. Both columns contain a list of categories: 'Inventario completo', 'Lustro', 'Área temática', 'Medio', 'Orixe', 'Bloque', 'Subtipo', 'Xénero', and 'Sexo-autor'.

Fig. 2. Acceso directo ás listaxes máis comúns de frecuencias.

Ao premer nalgún deles, por exemplo os elementos gramaticais consonte o *Lustro*, accédese directamente aos resultados da procura, sen ter que formalizala no sistema de consultas. Visualízanse así as frecuencias de todos os elementos do corpus distribuídos por lustro, onde poderemos comprobar que, malia ser a preposición *de* o elemento máis común na totalidade do corpus, se atendemos á distribución por lustros, o sinal de puntuación coma ‘ , ’ presenta unha frecuencia absoluta e relativa maior nos catro primeiros lustros:



The screenshot shows the search results page for the CORGA corpus. The interface includes a search bar, filters, and a table of results. The table shows the frequency of grammatical elements across different lustros. The first two rows of the table are highlighted.

	Total (54.737.277 / 57.693)	1975-1979 (978.095 / 117)	1980-1984 (1.669.899 / 143)	1985-1989 (2.019.335 / 120)	1990-1994 (5.414.080 / 386)	1995-1999 (9.829.561 / 12.422)	2000-2004 (7.341.493 / 9479)	2005-2009 (9.536.629 / 13.877)	2010-2014 (10.565.828 / 14.936)	2015-2019
1 de Preposición	3.446.825 62.970 / 57.567	59.540 60.873 / 117	97.997 58.684 / 143	117.359 58.118 / 119	312.076 57.642 / 381	622.374 63.317 / 12.404	485.726 66.162 / 9474	617.318 64.731 / 13.837	670.424 63.452 / 14.898	443.137 6
2 , Signo de puntuación	3.132.228 57.223 / 56.985	65.648 67.118 / 117	108.207 64.799 / 143	126.308 62.549 / 119	331.084 61.152 / 372	558.274 56.795 / 12.291	423.956 57.748 / 9454	549.665 57.637 / 13.697	583.535 55.229 / 14.691	361.975 5

Fig. 3. Listaxe inicial da frecuencia de elementos distribuídos por lustro.

Para obter as frecuencias doutras combinacións non incluídas na relación de **Listaxes personalizadas**, o usuario debe acceder ao sistema de consultas e formalizar a procura en función dos seus intereses particulares tendo en conta as características que se refiren nesta guía no apartado [5.7. Inventario](#).

En **Buscas** accédese ao núcleo do sistema de consultas. Nel, a recuperación de datos encarréirase na pestana **Corpus** cara a dúas posibilidades:

- O corpus completo, para o cal hai que escoller a opción **Etiquetado automaticamente**
- O subcorpus empregado como adestramento, cuxa etiquetaxe se revisou á man, para o cal hai que seleccionar **Etiquetado manualmente**

A procura en calquera deles pode organizarse polos seguintes parámetros, combinables entre si como veremos máis adiante:

- Palabras ortográficas: *cancela, falar, disllo, nel, ao redor do...*
- Elementos gramaticais: *cancela, falar, dis* (incluíndo todos os casos con pronomes enclíticos nos que entra *dis*: *dilo, disllo, dísnolo* etc.), *el* (incluíndo as contraccións *nel* e *del*), *ao redor de* (incluíndo os casos de *ao redor desta, ao redor duns, ao redor das* etc.).
- Lemas: *cancela* (inclúe os casos de *cancela* e *cancelas*), *falar* (todas as formas do paradigma do verbo *falar*), *el* (todos os casos do pronome tónico de terceira, masculinos, femininos ou non binarios, singulares ou plurais, e mais as concorrencias da forma arcaica do artigo determinado).
- Hiperlemas: *ditar* (inclúe todas as formas dos lemas *ditar* e *dictar*), *nin* (todas as ocorrencias de *nin* e *nen*), *el* (inclúe todas as formas agrupadas baixo os lemas *el* e *il*).
- Clases de palabras: substantivo, por exemplo.
- Valores das subcategorías gramaticais aplicables en cada caso: grao, xénero e número no caso dos adxectivos, por exemplo.

Respecto á visualización dos resultados, as opcións que se ofrecen para a extracción dos datos son as seguintes: *Frecuencia simple, Frecuencia completa, Concordancias, Inventario, Coaparicións* e *Nómina*. Delas iremos falando ao longo desta guía.

Por outra banda, o sistema permite a recuperación de datos da totalidade do corpus ou ben do subcorpus virtual creado directamente pola persoa que realiza a consulta en función das escollas que esta fai sobre os distintos filtros que é posible aplicar:

- Orixe, bloque, xénero e subtipo
- Medio
- Período cronolóxico
- Área temática
- Autor
- Sexo do autor, Sexo do interlocutor
- Documento, Obra

A maiores, seleccionando na pestana **Buscar en**, pode decidirse que a consulta se aplique sobre todo o documento, opción que aparece marcada por defecto, ou ben sobre unha parte estrutural concreta, por exemplo nos *titulares* das noticias xornalísticas ou nas *acoutacións* das obras de teatro e guións de series televisivas.

Por último, o parámetro referido á **Sensibilidade** permite que na recuperación de información o sistema teña en conta as diferenzas debidas ao emprego de **acentos ortográficos** e **maiúsculas**, de xeito que o usuario decide se para a súa consulta é relevante a distinción entre formas con tiles e formas sen tiles e/ou grafadas en minúsculas ou maiúsculas ou, pola contra, desexa que nos resultados se ignoren esas diferenzas.

2. O corpus

Os parámetros que vimos de enumerar son combinables entre si e poden empregarse sobre o corpus completo –opción **Etiquetado automaticamente** da pestana **Corpus**– ou ben sobre o subcorpus cuxa etiquetaxe se revisou á man –opción **Etiquetado manualmente** da pestana **Corpus**–.

Dende a versión 3.0. a aplicación unifica os dous sistemas de consulta en liña dispoñibles anteriormente (o do *Corpus de Referencia do Galego Actual* [CORGA] e o do *Corpus de Referencia do Galego Actual etiquetado* [CORGAetq]) nunha única plataforma que dá cabida ás diferentes aproximacións de busca: consulta por palabras ortográficas, elementos gramaticais, clases de palabras, etiquetas ou lemas, tanto no subcorpus desambiguado manualmente (antigo CORGAetq e actual **Etiquetado manualmente**) coma no CORGA, o cal se etiqueta na súa totalidade dun xeito automático (**Etiquetado automaticamente**).

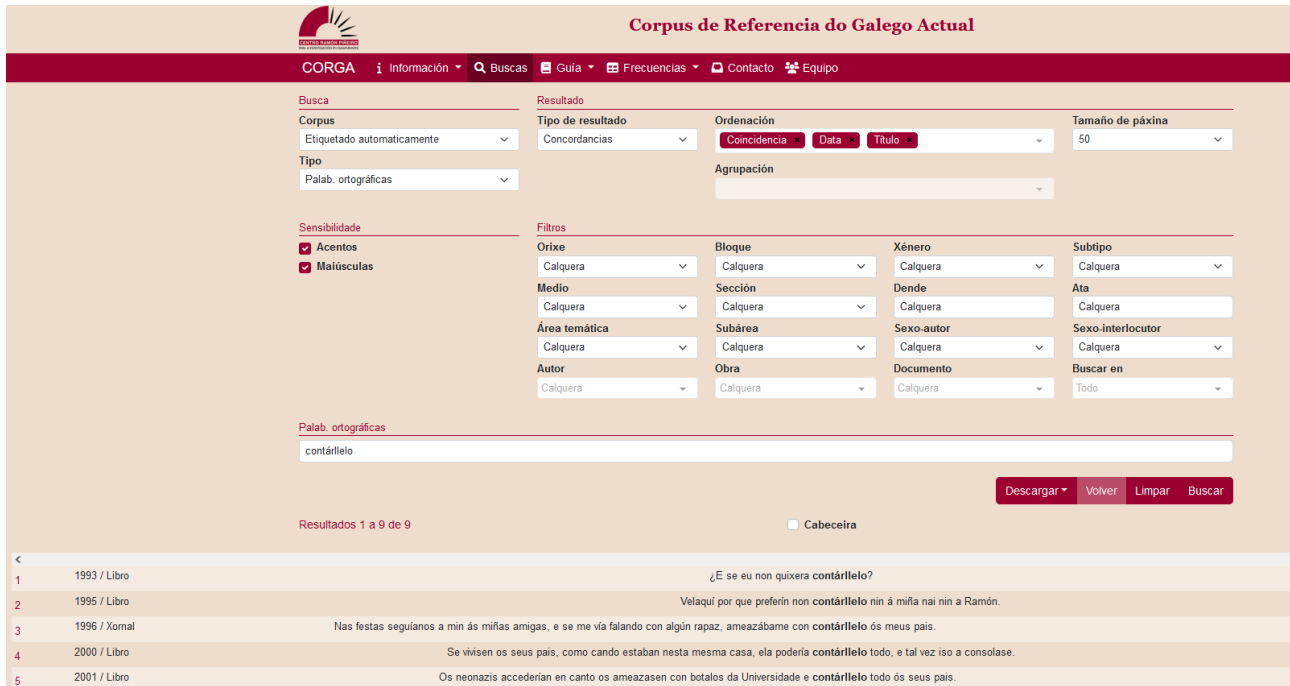
A opción **Etiquetado automaticamente** corresponde a unha versión do CORGA que presenta numerosas particularidades, entre as cales cabe destacar as seguintes:

- Acceso ao corpus sen necesidade de rexistrarse no sistema para poder consultalo.
- Etiquetaxe automática de todos os documentos, de xeito que son factibles as consultas por unha palabra completa, parte dunha palabra ou varias palabras ou partes, contiguas ou non, mais tamén poden realizarse consultas nas que estean implicados elementos gramaticais, clases de palabras, subcategorías gramaticais e lemas na totalidade dos textos incluídos.
- Inclusión de transcricións ortográficas procedentes de material sonoro radiofónico nas que se aliña o texto co son; na recuperación de información pódese reproducir o audio correspondente ao contexto da busca realizada.
- Clasificación dos documentos diferenciando entre *tipo de documento* (é dicir, se a súa orixe é escrita ou oral, se pertence ao bloque da ficción ou da non ficción, o xénero no que se encadra e o subtipo do que se trata) e *área temática* (aplicable só para os documentos da non ficción).
- Incremento do volume textual coa inclusión de novos documentos pertencentes aos últimos lustros.
- Inserción de documentos pertencentes a novos medios: *Internet* (integra polo momento blogs) e *audiovisual* (clasifica as transcricións e os guións de series televisivas).

O corpus **Etiquetado automaticamente** contén todos os documentos integrados no CORGA etiquetados dun xeito automático co *Etiquetador/Lematizador do Galego Actual (XIADA)*¹, desenvolvido conxuntamente polo Centro Ramón Piñeiro para a investigación en humanidades e o grupo COLE das universidades da Coruña e Vigo.

Como é lóxico, nos resultados da consulta por **Palabras ortográficas** mantéñense as maiúsculas e a ortografía convencional:

1 Este etiquetador, posuidor dunha taxa de acerto segundo datos do 2009 do 96 % (<https://doi.org/10.32766/cdl.30-31.37>), está dispoñible en liña no enderezo <http://corpus.cirp.gal/xiada> e pode utilizarse para etiquetar o texto que se lle proporcione, ben directamente ben a través de arquivo, sempre que non exceda de 100 secuencias. Por outra parte, no ano 2019 procedeuse á súa liberación e dende entón está dispoñible para a súa descarga en <https://github.com/crpih/xiada>.



The screenshot shows the 'Corpus de Referencia do Galego Actual' interface. The search criteria are set to 'Palabras ortográficas' and 'Etiquetado automáticamente'. The search term 'contárllelo' is entered in the search box. The results list shows five entries with their respective years and document types, along with the context of the word's use in each.

Year	Document Type	Context
1993	Libro	¿E se eu non quixera contárllelo?
1995	Libro	Velaquí por que preferín non contárllelo nin á miña nai nin a Ramón.
1996	Xornal	Nas festas seguíanos a min ás miñas amigas, e se me vía falando con algún rapaz, ameazábame con contárllelo ós meus pais.
2000	Libro	Se vivisen os seus pais, como cando estaban nesta mesma casa, ela podería contárllelo todo, e tal vez iso a consolase.
2001	Libro	Os neonazis accederían en canto os ameazasen con botalos da Universidade e contárllelo todo ós seus pais.

Fig. 4. Pantalla de concordancias da consulta por palabras ortográficas.

Nas buscas por **Palabras ortográficas**, inda que se teña seleccionado o corpus **Etiquetado automáticamente**, só poden consultarse formas ortográficas, parciais ou completas (*dicirlllo*, *noutras...* contan coma unha palabra). Con esta opción unicamente é posible a consulta por palabra ortográfica; é dicir, pode buscarse unha palabra completa, parte dunha palabra ou varias palabras ou partes destas, mais non é posible empregar lemas ou etiquetas. Manter esta modalidade dentro do CORGA garante a fiabilidade das consultas que se realicen co parámetro **Palabras ortográficas** (dado que non existe análise automática, non hai posibilidade de erro nos resultados obtidos), e permite a recuperación de información exacta cando a diferenza se establece na representación ortográfica e non nos elementos gramaticais (casos de ambigüidade segmental). Pensemos por exemplo no encontro da conxunción comparativa *ca* co artigo determinado. Se tokenizásemos todo o corpus e prescindíssemos da palabra ortográfica, non se poderían discriminar as ocorrencias en que esta combinación aparece contracta (*có*, *cá*, *cós*, *cás*) daquelas outras ocorrencias en que aparece sen contraer (*ca o*, *ca a*, *ca os*, *ca as*).

Agora ben, na consulta por **Elementos gramaticais** os textos, *grosso modo*, segméntanse nos seus elementos gramaticais constituíntes (*dicirlllo* por exemplo desagregase en *dicir*, *lle* e mais *o*) e cada unha desas unidades gramaticais delimitadas recibe unha caracterización morfosintáctica (a categoría gramatical e mais os valores aplicables en cada caso segundo o contexto no que se localiza) e remítese ao lema que lle corresponde. Todo o proceso é automático e non hai, en consecuencia, supervisión lingüística sobre o resultado. Este debe ser o corpus seleccionado cando na consulta se queiran especificar categorías gramaticais (obter as ocorrencias do adverbio *mañá*, pero non os casos substantivos por exemplo), agrupar en lemas as distintas formas dun paradigma (ocorrencias de todas as formas do verbo *ir*), localizar as formas que comparten algún trazo morfosintáctico determinado (todos os casos da segunda persoa do plural do infinitivo conxugado de calquera verbo) ou a combinación de calquera das variantes anteriores (todas as ocorrencias das formas plurais do verbo *ir* seguidas de infinitivo).

Pola súa banda, o corpus **Etiquetado manualmente** correspóndese co corpus de adestramento do xénero xornalístico e do de ficción para o *Etiquetador/Lematizador do Galego Actual* (XIADA).

A razón de ser deste subcorpus, etiquetado automaticamente e revisado á man por unha lingüista, é servir de adestramento para o etiquetador; porén, debido ao seu tamaño e á minuciosidade da etiquetaxe practicada, estimamos que pode ser de utilidade, fundamentalmente, para o estudo de aspectos gramaticais.

Á pequena modificación sobre a marcade dos segmentos que estaban nunha lingua diferente do galego² e mais a nova clasificación textual, características engadidas na versión 3.0, xunto coa inclusión da etiqueta específica para os termos de nomenclatura científica binomial (*Caenorhabditis elegans* e *Tursiops Truncatus*) e a reanotación como unidades multipalabras noutros casos (*tea de araña*, *de boca en boca*, *efecto invernadoiro*, *en base a* etc.), agregadas na versión 4.0, a presente versión incorpora, entre outras, a corrección dalgúns erros de etiquetaxe, a inclusión do parámetro *Sexo do interlocutor* ou *Obra* nos filtros e continúa a reanotación como unidades multipalabras noutros casos (*base de datos*, *en resumo*, *tarifa plana*, *ácido fólico*, *por exemplo* etc.).

As posibilidades de consulta no corpus **Etiquetado manualmente** son practicamente as mesmas ca no caso do **Etiquetado automaticamente**. As diferenzas entre eles estriban, sobre todo, no tamaño do subcorpus, inferior a un millón de elementos gramaticais, moito menor entón, e na modalidade da etiquetaxe, pois aquí existiu supervisión lingüística. Con todo, cómpre sinalar que as consultas segundo o sexo do autor e o sexo do interlocutor no corpus **Etiquetado manualmente** non devolverán resultados, mentres que as realizadas por hiperlema só devolverán os casos coincidentes co lema, non os doutros lemas irmáns. Isto débese a que, para o sexo do autor e o interlocutor, por ser posterior o seu desenvolvemento á construción do subcorpus, entendemos que carecía de relevancia para o adestramento e polo tanto non implementamos eses atributos nos metadatos dos arquivos que o compoñen, mentres que a existencia de ambigüidades formais para unha posible atribución automática do hiperlema impide que esta se realice sen supervisión, tafera que non se afrontou polo momento.

3. O sistema de buscas

O sistema de buscas é a cerna da aplicación, o lugar dende o que se accede e recupera a información contida no corpus e, así mesmo, o único lugar dinámico da aplicación, pois a información que aí vai aparecer determínaa o usuario mediante as escollas que realiza nos distintos parámetros de procura. Accédese a el premendo na pestana **Buscas** da liña superior da pantalla, xusto onde aparece unha lupa como símbolo da busca.

Ao premer na pestana **Buscas** atopamos unha serie de seccións que organizan a consulta e nas que a persoa usuaria pode precisar os parámetros da súa procura. Este é o seu aspecto por defecto:

2 Nas versións anteriores prescindíase por completo dos segmentos pertencentes a linguas diferentes do galego; con todo, dende a versión 3.0 mantéñense cunha etiqueta de *outra_lingua* para que o sistema non dea como contiguas na recuperación de información a forma anterior e posterior ao fragmento identificado con *outra_lingua*.



The screenshot shows the initial search interface of the CORGA (Corpus de Referencia do Galego Actual) website. The interface is organized into several sections:

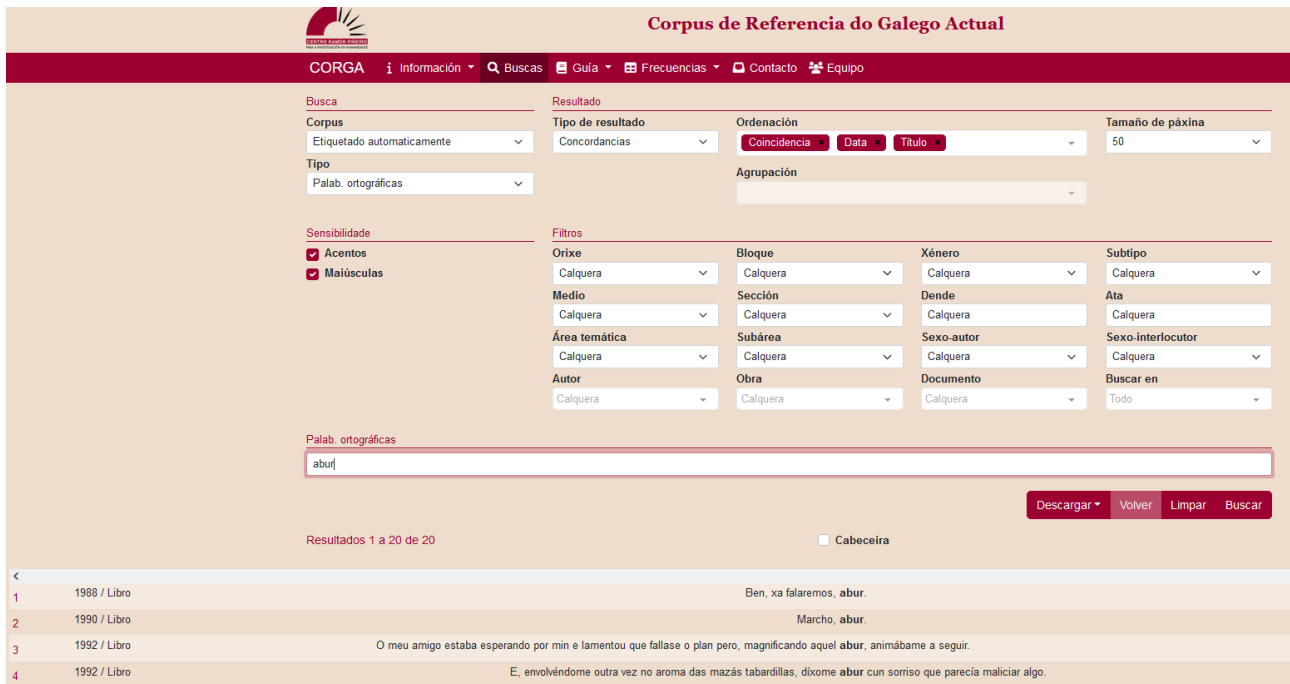
- Busca (Search):** Includes a 'Corpus' dropdown menu set to 'Etiquetado automaticamente', a 'Tipo' dropdown set to 'Palab. ortográficas', and a 'Sensibilidade' section with checkboxes for 'Acentos' and 'Maiúsculas'.
- Resultado (Results):** Features a 'Tipo de resultado' dropdown set to 'Concordancias', an 'Ordenación' section with buttons for 'Coincidencia', 'Data', and 'Título', and a 'Tamaño de páxina' dropdown set to '50'.
- Filtros (Filters):** A grid of dropdown menus for various filters: 'Orixe', 'Medio', 'Área temática', 'Autor', 'Bloque', 'Sección', 'Subárea', 'Obra', 'Xénero', 'Dende', 'Sexo-autor', 'Documento', 'Subtipo', 'Ata', 'Sexo-interlocutor', and 'Buscar en'.
- Palab. ortográficas (Orthographic words):** A text input field with the placeholder 'Cinco palabras máximo'.
- Buttons:** 'Volver', 'Limpar', and 'Buscar' buttons are located at the bottom right.

Fig. 5. Pantalla inicial de consulta.

Como se observa na figura 5, as seccións nas que se organiza a consulta son **Busca**, **Resultado**, **Sensibilidade**, **Filtros** e **Palabras ortográficas**.

Na sección **Busca** escóllese o corpus sobre o que se quere realizar a consulta (**Etiquetado automaticamente** ou **Etiquetado manualmente**) e o tipo de procura que se desexa realizar, podendo escoller se a busca se organiza por **Palabras ortográficas** (*dicirillo*) ou por **Elementos gramaticais** (*dicir, lle, o*), e se interesa a proximidade con respecto a outra palabra ortográfica ou elemento gramatical, en cuxo caso hai que seleccionar respectivamente **Palabras ortográficas próximas** ou **Elementos gramaticais próximos**.

A procura máis sinxela é por **Palabras ortográficas**, para o cal se completa coas formas ortográficas desexadas na caixiña habilitada, podendo introducir ata cinco palabras. Se escribimos no campo textual a palabra *abur* e deixamos nas demais seccións os valores que aparecen por defecto, despois de premer en **Buscar**, observamos que o sistema nos informa de que esa palabra aparece 20 veces e amósanos todos os casos no formato de concordancias. A opción que aparece por defecto no **Tipo** da sección **Resultado** é **Concordancias**, ou sexa, os casos concretos nos que aparece a forma obxecto da busca no formato KWIC (*key word in context*).



The screenshot shows the CORGA search interface. At the top, there is a navigation bar with 'CORGA' and various menu items like 'Información', 'Buscas', 'Guía', 'Frecuencias', 'Contacto', and 'Equipo'. Below this, there are several filter sections: 'Busca' (search criteria), 'Resultado' (result type and ordering), 'Sensibilidade' (sensitivity), and 'Filtros' (filters). The search term 'abur' is entered in the search box. The results section shows four entries, each with a number, year, and document type, followed by a snippet of text containing the word 'abur' in bold and highlighted in black.

Resultado	Orixe	Bloque	Xénero	Subtipo
1	1988 / Libro	Calquera	Calquera	Calquera
2	1990 / Libro	Calquera	Calquera	Calquera
3	1992 / Libro	Calquera	Calquera	Calquera
4	1992 / Libro	Calquera	Calquera	Calquera

Fig. 6. Concordancias da forma *abur*.

Como se pode observar, cada ocorrencia da palabra obxecto da busca aparece nunha liña independente, centrada e destacada en negra co contexto inmediato, tanto anterior coma posterior. A información que contén cada liña, de dereita a esquerda, é a seguinte:

- A concordancia da forma obxecto da busca, aparecendo esta centrada e destacada en negra. Para facer máis cómoda a lectura, as marcas de codificación non aparecen no texto³, pero sinálase a súa existencia mediante o destacado do texto en amarelo, e só cando se pasa co punteiro do rato por enriba emerxe un cadro de texto no que se indica a que corresponde a marcaxe (alongamento, palabra cortada, outra lingua etc.). Para máis información sobre as distintas etiquetas utilizadas, véxase a relación de etiquetas empregadas na codificación do corpus contidas no ítem *Etiquetas de codificación* baixo a pestana **Guía**.
- Referencia do ano e medio ao que corresponde a concordancia da ocorrencia. Pasando o punteiro do rato por riba desta zona emerxe un cadro de texto no que se recollen todos os metadatos correspondentes á cabeceira do documento da concordancia concreta.
- Punteiro dunha frecha no caso de que a ocorrencia se localice nunha transcripción do oral (resultado 11 da procura realizada). Ao premer na frecha actívase a reprodución do son co que está aliñado o texto da concordancia.
- Número de orde da concordancia. Se se preme nel accédese a unha ampliación do contexto, tanto da secuencia na que se atopa o exemplo, como das dúas secuencias anteriores e posteriores, se as houber. No contexto incorpórase tamén a información relativa ao *interlocutor/falante* ao que se remite cada secuencia, se o documento no que se documenta a ocorrencia é unha transcripción, unha obra de teatro, un guión ou unha entrevista. Inclúese, así mesmo, en idéntico formato ao de *interlocutor*, a información de se nun texto dramático a secuencia corresponde a unha *acoutación*. Ademais, todas as secuencias do contexto dunha ocorrencia que se localice nunha transcripción poden reproducir o son dende alí premendo no punteiro da frecha que aparece en cada caso. Por último, toda a información anterior aparece

3 A excepción constitúena as etiquetas correspondentes a subíndice, superíndice, táboa e fórmula.

encabezada cos metadatos pertinentes do documento correspondentes á concordancia concreta (título, autor, sexo do autor, editorial –ou url se se trata dun texto que só se rexistra en Internet–, ano, medio, orixe, bloque, xénero, subtipo etc.).



Palab. ortográficas
abur

Volver Limpar Buscar

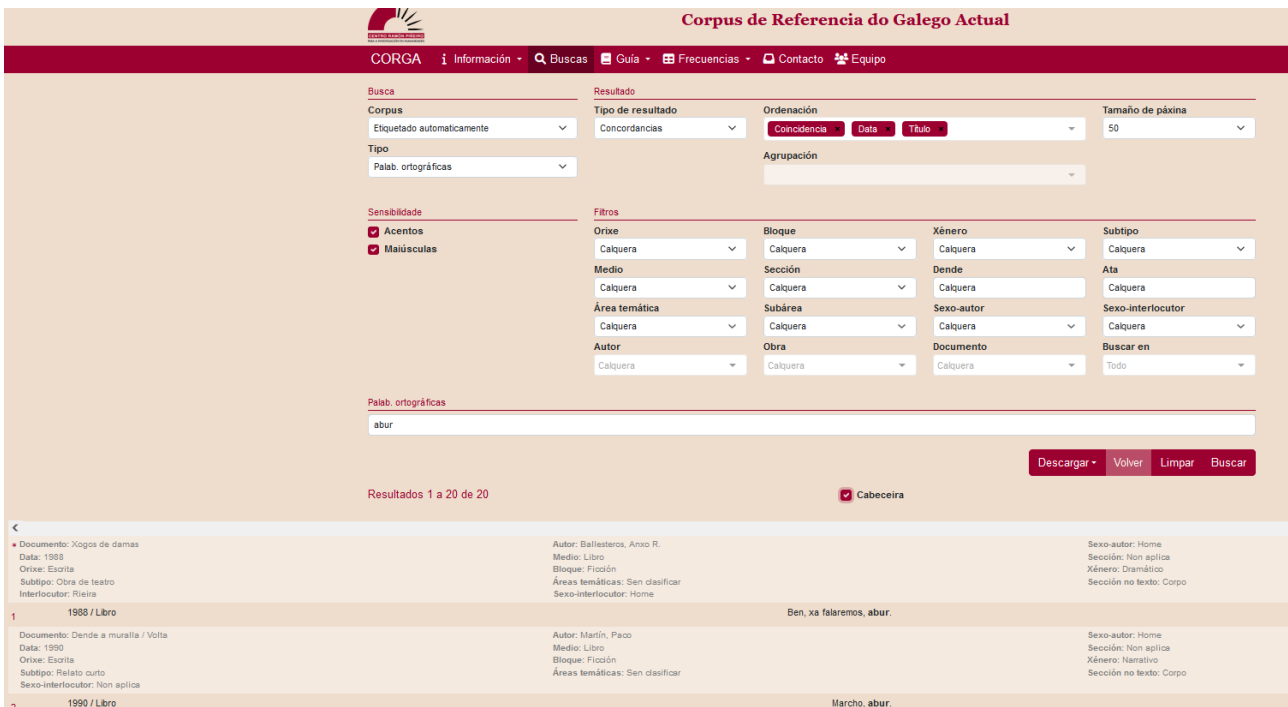
Contexto do exemplo 11 da listaxe anterior.

Documento: Pensando en ti. 18/06/1997	Sexo-autor: Non aplica	Data: 1997
Medio: Audiovisual	Sección: Non aplica	Orixe: Oral
Bloque: Non aplica	Xénero: Non aplica	Subtipo: Variedades
Áreas temáticas: Sen clasificar	Interlocutor: Xosé Antonio	Sexo-interlocutor: Home

- ▶ **Xosé Antonio:** non levamos tempo <pausa/> **eso** pois alégrome de escoitarle outra vez <pausa/> e saudíños para todos
- ▶ **Xosé Antonio:** e nada máis <pausa/> que me vou durmir <pausa/> que é a hora xusta para min
- ▶ **Xosé Antonio:** hasta outro momentíño <pausa/> unha aperta <pausa/> **abur**
- ▶ **Marcial Mouzo:** unha aperta Xosé Antonio ata outro momento
- ▶ **Marcial Mouzo:** con ben por aí e xa sabes onde estamos cada **madrugada** <pausa/> cando queiras darnos unha chamadiña

Fig. 7. Contexto da ocorrencia da forma *abur* nunha transcripción.

É posible tamén ver os datos completos que localizan cada ocorrencia xa na pantalla de resultados, sen necesidade de acudir ao contexto, con só seleccionar a opción **Cabeceira** que aparece inmediatamente antes dos resultados.



Corpus de Referencia do Galego Actual

CORGA i Información Búscas Guía Frecuencias Contacto Equipo

Busca
Corpus
Etiquetado automaticamente
Tipo
Palab. ortográficas

Resultado
Tipo de resultado
Concordancias
Ordenación
Concordancia Data Título
Tamaño de páxina
50
Agrupación

Sensibilidade
 Acentos
 Maiúsculas

Filtros
Orixe
Calquera
Medio
Calquera
Área temática
Calquera
Autor
Calquera
Bloque
Calquera
Sección
Calquera
Subárea
Calquera
Obra
Calquera
Xénero
Calquera
Dende
Calquera
Sexo-autor
Calquera
Documento
Calquera
Subtipo
Calquera
Ata
Calquera
Sexo-interlocutor
Calquera
Buscar en
Todo

Palab. ortográficas
abur

Descargar Volver Limpar Buscar

Resultados 1 a 20 de 20 Cabeceira

1	1988 / Libro	Ben, xa falaremos, abur.
2	1990 / Libro	Marcho, abur.

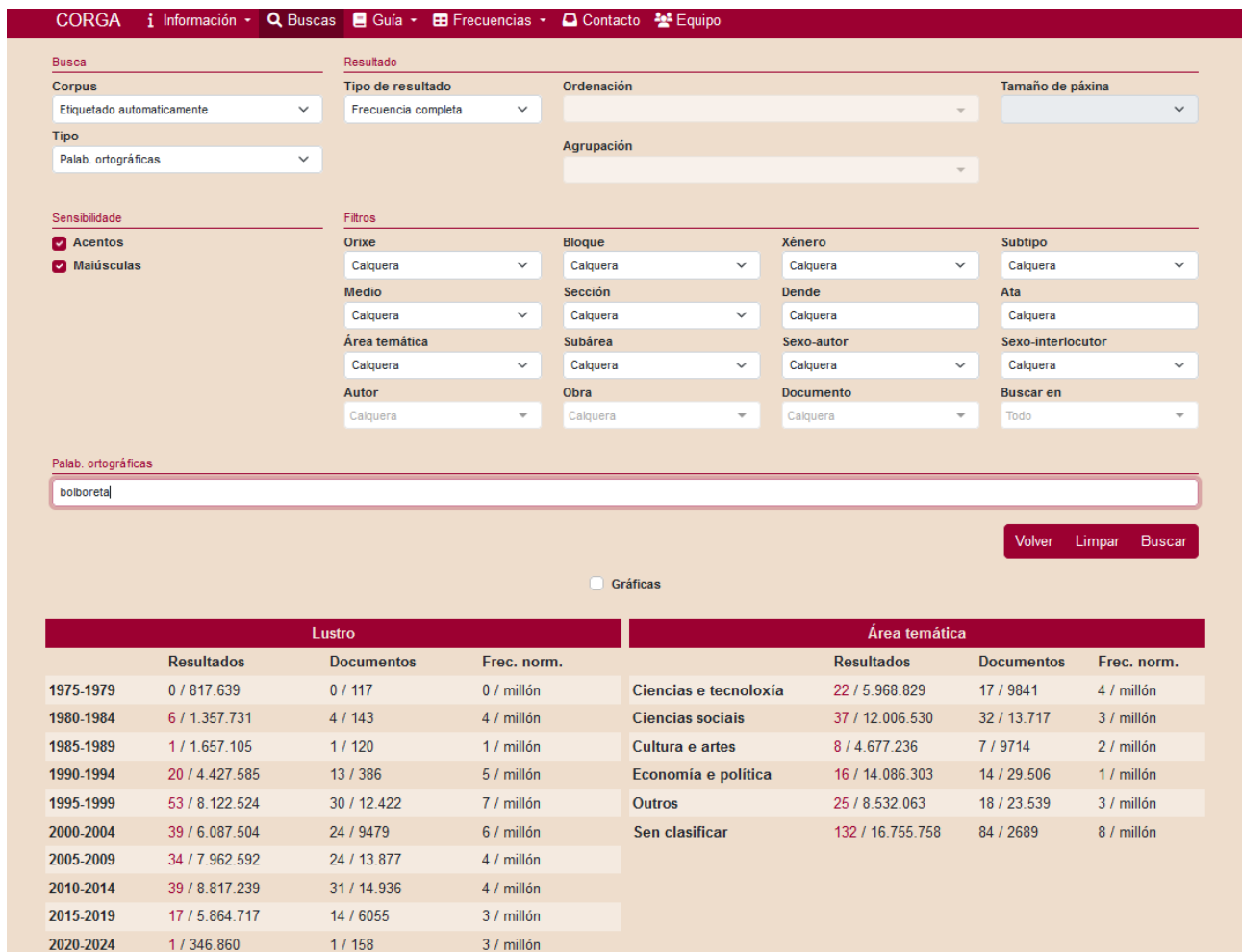
Fig. 8. Concordancias da forma *abur* cos datos da súa localización visibles.

Entre as opcións que ofrece a pantalla de resultados coas concordancias da unidade obxecto da busca están **Volver**, para regresar á páxina de resultados dende a pantalla do contexto ampliado, e **Descargar**, se se desexa exportar o resultado para examinar os datos con máis detemento.

Para calquera consulta que se efectúe, a aplicación devolve a **Frecuencia simple**, a **Frecuencia completa** e as **Concordancias** do resultado da procura (para as modalidades **Inventario**, **Coaparicións** e **Nómina** remitimos á epígrafe respectiva). A opción que figura por defecto no campo

Tipo é a de **Concordancias**, polo que se o usuario desexa coñecer o número de ocorrencias por millón de palabras debe modificar o valor seleccionado para **Frecuencia simple**, sen que sexa preciso escribir novamente nada no campo **Texto**; se pola contra desexa coñecer a **Frecuencia completa** deberá marcar estoutro valor.

Os datos, obviamente, son os mesmos en ambos os dous casos, na frecuencia simple e na completa, mais nesta última amósanse distribuídos segundo os parámetros de estruturación dos documentos no corpus: *lustro*, *área temática*, *medio*, *orixe*, *bloque*, *xénero*, *subtipo*, *sexo do autor* e *sexo do interlocutor*, obtendo deste xeito unha panorámica da distribución dunha forma dada con respecto aos parámetros presentes na clasificación dos textos integrados no corpus. Recollemos a xeito de exemplo, fragmentadas nas dúas imaxes seguintes por cuestións de representación, as frecuencias completas que ofrece o sistema de recuperación de información para a palabra ortográfica *bolboreta*:



The screenshot shows the CORGA search interface. The search term is 'bolboreta'. The interface includes various filters for Corpus, Tipo, Sensibilidade, and Filtros (Orixe, Medio, Área temática, Autor, Bloque, Sección, Subárea, Obra, Xénero, Dende, Sexo-autor, Documento, Subtipo, Ata, Sexo-interlocutor, and Buscar en). The results are displayed in a table with two main sections: Lustro and Área temática.

	Lustro			Área temática			
	Resultados	Documentos	Frec. norm.	Resultados	Documentos	Frec. norm.	
1975-1979	0 / 817.639	0 / 117	0 / millón	Ciencias e tecnoloxía	22 / 5.968.829	17 / 9841	4 / millón
1980-1984	6 / 1.357.731	4 / 143	4 / millón	Ciencias sociais	37 / 12.006.530	32 / 13.717	3 / millón
1985-1989	1 / 1.657.105	1 / 120	1 / millón	Cultura e artes	8 / 4.677.236	7 / 9714	2 / millón
1990-1994	20 / 4.427.585	13 / 386	5 / millón	Economía e política	16 / 14.086.303	14 / 29.506	1 / millón
1995-1999	53 / 8.122.524	30 / 12.422	7 / millón	Outros	25 / 8.532.063	18 / 23.539	3 / millón
2000-2004	39 / 6.087.504	24 / 9479	6 / millón	Sen clasificar	132 / 16.755.758	84 / 2689	8 / millón
2005-2009	34 / 7.962.592	24 / 13.877	4 / millón				
2010-2014	39 / 8.817.239	31 / 14.936	4 / millón				
2015-2019	17 / 5.864.717	14 / 6055	3 / millón				
2020-2024	1 / 346.860	1 / 158	3 / millón				

Fig. 9. Parte inicial da pantalla de frecuencias de *bolboreta*.

Medio			Orixe			
	Resultados	Documentos	Frec. norm.	Resultados	Documentos	Frec. norm.
Audiovisual	0 / 814.255	0 / 157	0 / millón	Escrita 210 / 44.771.915	142 / 57.554	5 / millón
Internet	1 / 92.593	1 / 207	11 / millón	Oral 0 / 689.581	0 / 139	0 / millón
Libro	148 / 24.667.056	96 / 3297	6 / millón			
Revista	13 / 6.314.101	12 / 9718	2 / millón			
Xornal	48 / 13.573.491	33 / 44.314	4 / millón			

Bloque			Subtipo			
	Resultados	Documentos	Frec. norm.	Resultados	Documentos	Frec. norm.
Ficción	134 / 16.160.488	86 / 2610	8 / millón	Artigo científico 0 / 321.849	0 / 52	0 / millón
Non aplica	0 / 689.581	0 / 139	0 / millón	Blog 1 / 92.593	1 / 207	11 / millón
Non ficción	76 / 28.611.427	56 / 54.944	3 / millón	Conferencia 0 / 10.961	0 / 6	0 / millón
				Divulgación 14 / 7.524.991	10 / 666	2 / millón
				Entrevista 0 / 39.393	0 / 12	0 / millón
				Guión 0 / 148.568	0 / 20	0 / millón
				Informativo 0 / 115.684	0 / 23	0 / millón
				Libro de texto 0 / 512.409	0 / 7	0 / millón
				Memoria 0 / 477.775	0 / 10	0 / millón
				Novela 76 / 11.352.312	41 / 240	7 / millón
				Obra de teatro 20 / 1.368.318	14 / 193	15 / millón
				Programa cultural 0 / 104.057	0 / 13	0 / millón
				Publicidade 0 / 8253	0 / 29	0 / millón
				Relato curto 38 / 3.291.290	31 / 2157	12 / millón
				Revista 13 / 6.108.319	12 / 9688	2 / millón
				Tertulia 0 / 174.941	0 / 26	0 / millón
				Variedades 0 / 236.292	0 / 30	0 / millón
				Xornal 48 / 13.573.491	33 / 44.314	4 / millón

Xénero			
	Resultados	Documentos	Frec. norm.
Dramático	20 / 1.516.886	14 / 213	13 / millón
Ensaístico	14 / 8.837.024	10 / 735	2 / millón
Narrativo	114 / 14.643.602	72 / 2397	8 / millón
Non aplica	0 / 689.581	0 / 139	0 / millón
Xornalístico	62 / 19.774.403	46 / 54.209	3 / millón

Sexo-autor			Sexo-interlocutor			
	Resultados	Documentos	Frec. norm.	Resultados	Documentos	Frec. norm.
Ambos	1 / 889.533	1 / 232	1 / millón	Ambos 0 / 2342	0 / 51	0 / millón
Descoñecido	28 / 11.201.866	20 / 40.062	2 / millón	Descoñecido 0 / 61.508	0 / 472	0 / millón
Home	156 / 26.161.303	102 / 12.388	6 / millón	Home 4 / 2.560.664	4 / 1495	2 / millón
Muller	25 / 6.517.049	19 / 4864	4 / millón	Muller 7 / 958.928	4 / 909	7 / millón
Non aplica	0 / 691.745	0 / 147	0 / millón	Non aplica 196 / 41.786.728	133 / 57.542	5 / millón
				Non binario 0 / 1868	0 / 5	0 / millón
				Non pertinente 3 / 89.458	3 / 75	34 / millón

Fig. 10. Parte final da pantalla de frecuencias de *bolboreta*.

A consulta por **Frecuencia completa** ofrece, así mesmo, a posibilidade de presentar a información mediante gráficas. Para iso só hai que seleccionar a opción **Gráficas** que precede os resultados das frecuencias completas, sen necesidade de premer de novo en **Buscar**. As gráficas simplifican a información e preséntana dunha forma dinámica e atractiva para o usuario, quen obtén deste xeito, dunha maneira rápida e directa, unha idea xeral da distribución dos elementos obxecto da procura segundo os diversos parámetros que interveñen na clasificación textual.

O sistema do CORGA, ademais de permitir visualizar a información das frecuencias completas representada en gráficas, permite navegar entre estas e as ocorrencias dun xeito paralelo, posto que premendo nunha das variables incluída nalgunha das gráficas pódese acceder directamente ás súas ocorrencias, e para regresar ao resultado inicial das gráficas basta con premer na pestana **Volver**.

Seguindo co exemplo de arriba, a información gráfica facilítanos comprobar cunha simple ollada que para a palabra *bolboreta* destaca a gráfica circular que distribúe os seus usos atendendo á orixe do texto, oral ou escrita, pois conséntase no corpus o seu emprego exclusivo en textos escritos:



Fig. 11. Pantalla inicial da información gráfica sobre a palabra *bolboreta*.

As gráficas relativas á distribución temporal e ás áreas temáticas representan a frecuencia normalizada, a cal, dada a baixa frecuencia dos fenómenos lingüísticos, adoita expresarse en porcentaxes por millón. Isto pode provocar nalgunha busca determinada que a gráfica correspondente non amose resultados para un ou varios rangos, polo que se aconsella sempre corroborar os datos cos que ofrecen as frecuencias completas normais, pois pode darse que si se rexistren datos para un dos rangos nunha procura concreta, mais que estes non se amosen nas gráficas por seren inferiores a 1/millón.

A escolla das opcións **Concordancias**, **Inventario**, **Coaparicións** ou **Nómina** no bloque **Tipo** para a visualización dos resultados da consulta determina a activación dos campos **Ordenación** e **Tamaño de páxina**, que se manteñen desactivados para os valores **Frecuencia simple** e **Frecuencia completa**. É dicir, só se está marcada unha das alternativas diferentes a **Frecuencia**, simple ou completa, o usuario pode decidir cantos resultados quere ver por páxina e de que xeito quere que se ordenen. As posibilidades de ordenación, combinables entre si, son as seguintes: *área temática*, *coincidencia*, *data*, *título*, *etiqueta*, *lema*, *medio*, *palabra anterior*, *palabra posterior*, *segunda palabra anterior* e *segunda palabra posterior*. Por defecto, o criterio consonte o cal se ordenan os

resultados das buscas é *coincidencia*, mais se nela non hai variacións ortográficas actúa a *data* como segundo parámetro.

Dende a versión 3.1 o usuario dispón, á parte das anteriores opcións de ordenación, da posibilidade de organizar os resultados polo número do compoñente, a etiqueta ou o lema que determine, sempre en función do número de elementos que procure dentro dos 5 que permite o sistema, pois as alternativas ampliáanse cos seguintes valores: *elemento 2, elemento 3, elemento 4 e elemento 5; etiqueta 2, etiqueta 3, etiqueta 4 e etiqueta 5; e lema 2, lema 3, lema 4 e lema 5*.

4. A descarga dos resultados

A aplicación ofrece a posibilidade de gardar a procura realizada e mais os resultados obtidos no botón **Descargar**. Con esta acción o usuario pode gardar o resultado no dispositivo que elixa, por defecto a carpeta de *descargas* do equipo dende o que se realiza a consulta.

Para descargar os resultados dunha procura ou un arquivo da listaxe de frecuencias (este último un .zip con varios ficheiros dentro) dispón de dúas posibilidades:

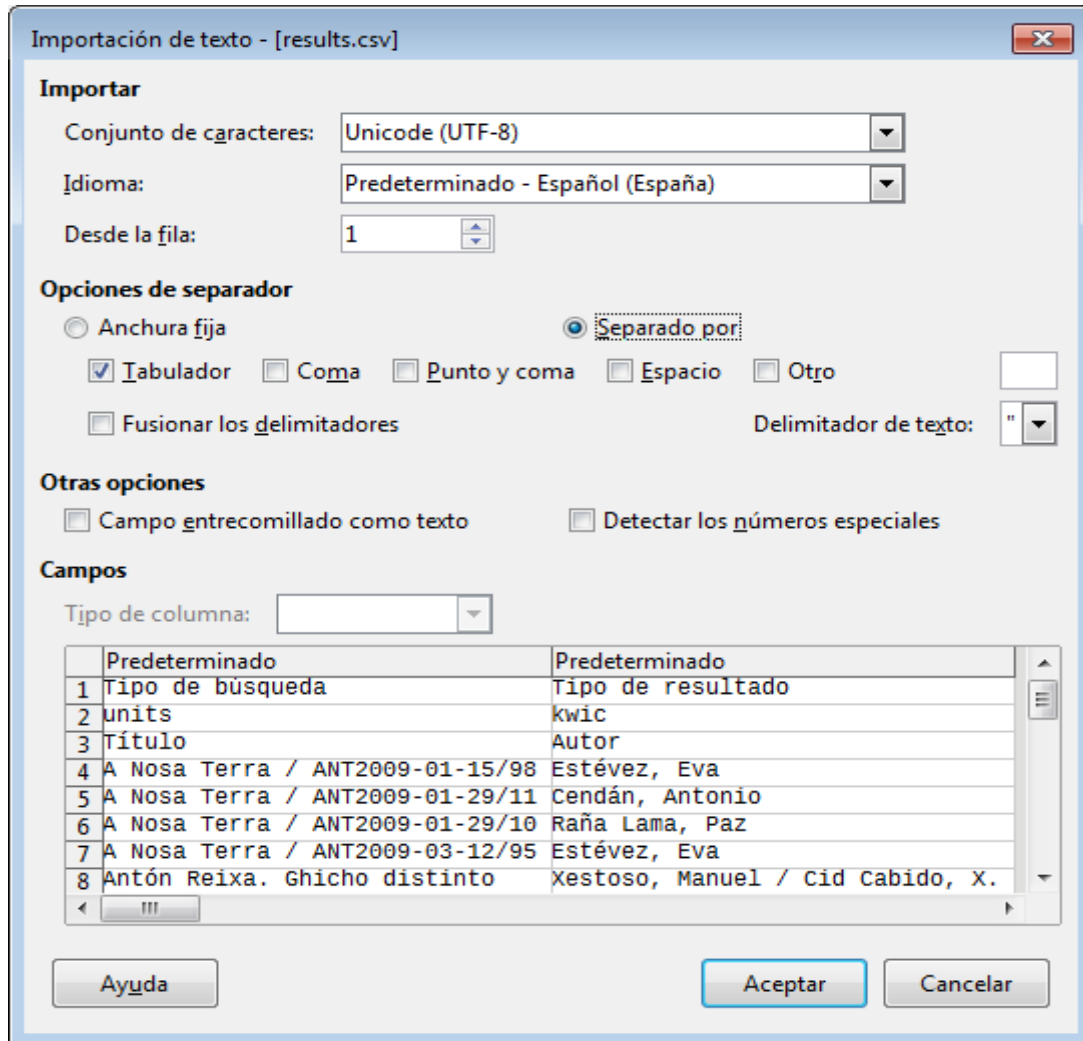
- 1) .csv
- 2) .xlsx

Se se descarga o arquivo .xlsx, con só premer sobre el xa se abrirá, independentemente de que se teña instalado Microsoft Excel ou LibreOffice.

Pola contra, o proceso complícase lixeiramente se se descarga o arquivo .csv, xa que se pode abrir con:

- 1) LibreOffice ou OpenOffice. Neste caso aparecerá un menú para escoller cal debe ser o carácter de separación (debe seleccionarse tabulador) e mais o sistema de codificación (debe seleccionarse UTF-8). Véxase a figura 12.

- 2) Microsoft Excel. Cómpre gardar o arquivo de resultados en vez de abri-lo directamente, logo executar o programa Microsoft Excel e seleccionar "Datos / Dende texto". Despois de elixir o arquivo descargado, no cadro ou cadros de diálogo que van aparecendo debe seleccionarse a codificación de caracteres UTF-8 e indicar que o carácter de separación de campos é o tabulador.



Importación de texto - [results.csv]

Importar

Conjunto de caracteres: Unicode (UTF-8)

Idioma: Predeterminado - Español (España)

Desde la fila: 1

Opciones de separador

Anchura fija Separado por

Tabulador Coma Punto y coma Espacio Otro

Fusionar los delimitadores Delimitador de texto: "

Otras opciones

Campo entrecomillado como texto Detectar los números especiales

Campos

Tipo de columna:

	Predeterminado	Predeterminado
1	Tipo de búsqueda	Tipo de resultado
2	units	kwic
3	Título	Autor
4	A Nosa Terra / ANT2009-01-15/98	Estévez, Eva
5	A Nosa Terra / ANT2009-01-29/11	Cendán, Antonio
6	A Nosa Terra / ANT2009-01-29/10	Raña Lama, Paz
7	A Nosa Terra / ANT2009-03-12/95	Estévez, Eva
8	Antón Reixa. Ghicho distinto	Xestoso, Manuel / Cid Cabido, X.

Ayuda Aceptar Cancelar

Fig. 12. Pantalla de importación dos resultados para o formato .csv.

Na parte inferior da imaxe, as dúas primeiras filas recollen a información referente aos parámetros da busca que realizamos; na fila 3 aparece o encabezamento de a que corresponde cada columna e a partir da fila 4 amósanse os resultados.

Nas **concordancias**, o número de resultados que se importa está en consonancia, en primeiro lugar, coa opción que se selecciona no **Tamaño de páxina** para ver o resultado da consulta: 50, 100, 500, 1000 ou 5000; e, en segundo lugar, coa páxina de resultados na que se atopa a aplicación no momento de solicitar a descarga, de xeito que se estamos na páxina 2 dos resultados, e o tamaño da páxina elixido é de 50, exportaremos do 51 ao 100. En cambio, para a modalidade **Inventario** o sistema ofrece tamén a opción de descargar todos os resultados, coas mesmas alternativas en canto a formatos e características ca por páxina.

O formato de devolución en campos separados por tabuladores facilmente integrable nunha folla de cálculo ofrece, así mesmo, a posibilidade de prescindir xa dalgunha das columnas dende as opcións de importación; só é preciso seleccionar as columnas non desexadas e marcar *ocultar* no *Tipo de columna* do menú que amosa a imaxe anterior. Este formato permite ademais todas as vantaxes asociadas a unha folla de cálculo, facilitando incluso a ordenación por características non incluídas na aplicación.

5. Buscas simples e buscas avanzadas

5.1. Metacaracteres e operadores booleanos

Á riqueza dos tipos de busca posibles –**Palabras ortográficas** (*dicir*llo), **Elementos gramaticais** (*dicir*, *lle*, *o*), **Palabras ortográficas próximas** e **Elementos gramaticais próximos**– debe engadirse o emprego en calquera deles de metacaracteres e operadores booleanos que facilitan a recuperación de información. Os signos que se utilizan son o signo de peche de interrogación (?), o signo de peche de admiración (!), o asterisco (*) e a barra vertical (|). Todos eles poden empregarse en calquera dos campos dos distintos tipos de busca e escribíranse sen deixar espazos nin antes nin despois⁴. En función do seu cometido, podemos agrupalos por pares:

1) Substitúen caracteres (metacaracteres)

? substitúe un carácter: *de?de* devolve tanto os casos de *desde* coma *dende*; *?ol?oreta* devolve todas as ocorrencias de *bolboreta* e *volvoreta*.

* substitúe cero, un ou varios caracteres. Por exemplo, para ver que variantes de *ó redor* aparecen na construción numérica aproximativa cómpre indicar no campo de texto **redor* e así nos resultados aparecerán as formas *arredor*, *darredor*, *derredor*, *orredor*, *ó redor*, *ao redor*, *ó (meu/teu/seu/noso/voso) redor* ou *ao (meu/teu/seu/noso/voso) redor*. Non obstante, entre os resultados tamén saen formas que nada teñen que ver co que desexamos obter: *acredor*, *corredor*, *tredor* e *varredor*, pois todos eles rematan na secuencia *-redor*.

2) Condicionan a presenza/ausencia de resultados (operadores booleanos)

! equivale a NON. Emprégase para impedir que entre os resultados apareza algún que conteña a información que se especifica inmediatamente despois do signo de admiración de peche (!). Por exemplo, se no campo de texto escribimos **ante!ante!diante*, o sistema devolve as concordancias de todas as palabras ortográficas que rematan en *-ante*, menos as correspondentes a *ante* e *diante*.

| equivale a OU. Serve para introducir máis dunha alternativa no obxecto da consulta. O sistema devolverá as ocorrencias da opción que situamos en primeiro lugar, mais tamén os casos da opción que segue inmediatamente a barra vertical. Por exemplo, escribindo no campo de texto *ante|perante* obtemos todas as ocorrencias en que aparece calquera das dúas palabras.

Naturalmente, as catro expresións anteriores poden combinarse entre si, de xeito que chegan a realizarse buscas cun nivel de complexidade considerable. Por exemplo, supoñamos que nos interesa ver que palabras da familia léxica de *bolboreta* se recollen no corpus, mais non queremos que entre os resultados se inclúan os xa coñecidos *bolboreta* e *volvoreta*. Para iso, botando man dos comodíns, formulamos a consulta nos seguintes termos:

?ol?oret!bolboreta!volvoreta*

Traducido á lingua común, o anterior significa que:

- co signo de peche de interrogación, na posición que este ocupa, posibilitamos a aparición de *b*, *v* ou de calquera outro carácter;

⁴ Dado que ? e ! conflúen como comodíns nas expresións regulares e sinais de puntuación, cando se queiran usar nas buscas co valor de sinais de puntuación deberase precedelos da barra oblicua invertida (\): \? ou \!

- co asterisco final en *?ol?oret** damos cabida á aparición de calquera secuencia de caracteres completando a parte final da forma gráfica;
- finalmente, co signo de peche de admiración ante *bolboreta* e *volvoreta* impedimos que entre os resultados aparezan estas formas.

A idea de incluír un asterisco na parte final de *!bolboreta!volvoreta* para restrinxir nos resultados os correspondentes plurais débese meditar, pois bloquearíamos a aparición de posibles formas verbais con esta raíz.

A procura devólvenos os plurais *bolboretas*, *volvoretas*, os diminutivos *bolboretiñas* e *volvoretiña*, a unidade *bolboreteira* e mais as formas verbais *bolboretou*, *bolboreteaba* e *bolboreteaban* ou *volvoreteando*. Porén, tamén emerxen entre os resultados todas as ocorrencias de *Bolboreta* ou *Bolboretas*, porque estas outras formas non as excluímos do criterio de busca ao non ter en conta que o sistema é sensible á distinción entre maiúsculas e minúsculas.

5.2. Sensibilidade

A aplicación permite neutralizar na recuperación de información a sensibilidade aos acentos, parámetro de grande utilidade cando, coma no noso caso, o corpus contén documentos anteriores á publicación das primeiras normas ortográficas oficiais e os textos amosan unha variación acentual importante. Imaxinemos que queremos recuperar todas as ocorrencias do infinitivo *construír* para ver en que construcións aparece; pois ben, dado que neste caso importa máis a construción que as formas concretas do paradigma verbal, é de utilidade indicar que non desexamos que o sistema sexa sensible aos acentos, co que nos resultados obteremos tanto as concordancias de *construír* coma de *construir*.

Así mesmo, dende a versión 3.0 engádese a posibilidade de ter en conta a diferenza entre maiúsculas e minúsculas, co que se facilita o estudo da lexicalización de siglas, do emprego de marcadores discursivos a comezo de enunciado fronte ao seu uso en interior de secuencia etc. Naturalmente, nas consultas pode habilitarse ou deshabilitarse a recuperación de información atendendo á sensibilidade a acentos e maiúsculas e cruzar eses valores coas distintas expresións regulares, de xeito que se incrementa aínda máis a potencialidade da busca. Así, se quixésemos estudar as construcións nas que se atopa o verbo *desenvolver*, tras deshabilitar a sensibilidade aos acentos, poderíamos escribir no campo de texto *desenvol*!desenvolv?mento*!desenvoltura**, co que acadaríamos a recuperación de todas as formas verbais, con clíticos ou sen eles, pois o asterisco substitúe calquera cadea de caracteres e eliminamos as diferenzas debidas á acentuación.

Malia que a posibilidade de empregar expresións regulares e de neutralizar as distincións debidas a acentos gráficos e maiúsculas, ambas combinables entre si, incrementa substancialmente as potencialidades das buscas, as consultas por palabra ortográfica presentan serias limitacións, pois non son suficientes para obter datos gramaticais nos que se precisa distinguir entre clases de palabras, remitir as distintas formas dun paradigma a un lema ou abstraerse das formas concretas para buscar padróns sintácticos. Imaxinemos por exemplo a inxente cantidade de consultas que teríamos que realizar para poder extraer todas as concorrencias nas que están implicadas as formas verbais de *ter*. É verdade que con moita paciencia, en consultas sucesivas, deshabilitando a sensibilidade aos acentos para facilitar tamén a recuperación das formas verbais con enclíticos, poderíamos buscar *teñ** (para obter a primeira singular do presente de indicativo e todo o presente subxuntivo), *tiñ** (para o copretérito), *tiv** (para obter as formas do tema de pretérito), *tid** (para o participio) e *te** (para o presente de indicativo –fóra a primeira singular– e o infinitivo). Obteríamos entón, poñamos por caso, *teño*, *téñoos*, *teña*, *teñas*, *teñan*, *teñámola*, *teñades*, *tiña*, *tiñas*, *tiñamos* etc., mais tamén estarían entre os resultados os castellanismos *teñir* e *teñido*, o substantivo *tiña*, e

moitísimas outras formas coma *tiñoso, tese, tedescos, tenacidade, tenda, tiduo...* que teríamos que separar daqueles casos que corresponden formalmente co que pediamos, mais que non se corresponden co que en realidade pretendiamos.

Conscientes pois das limitacións que impoñen as consultas por forma ortográfica, no *Centro Ramón Piñeiro para a investigación en humanidades* trabállase tamén na mellora de varias ferramentas que permiten facer consultas moito máis avanzadas e propician dar un salto cualitativo nas posibilidades de busca, pois entendemos que o camiño a seguir é etiquetar o corpus; é dicir, remitir cada un dos elementos gramaticais a un lema e caracterizalo morfosintacticamente, de xeito que se especifique en cada caso a clase de palabra á que pertence e mais os valores gramaticais que sexan pertinentes. Así, no corpus para a palabra ortográfica *teño* ten que constar que é a primeira persoa do singular do presente de indicativo do verbo *ter*, e que *téñoos* é unha palabra ortográfica constituída por dous elementos gramaticais: a primeira persoa do singular do presente de indicativo do verbo *ter* e mais o pronome átono de terceira acusativo masculino plural cuxo lema é *o*.

A escolla da opción **Elementos gramaticais** na pestana do **Tipo** do bloque **Busca** activa unha nova sección na parte inferior da pantalla de consultas co acceso aos campos **Elemento gramatical**, **Etiqueta**, **Lema**, **Hiperlema** e **Unidade**.

5.3. Distinción entre palabra ortográfica, elemento gramatical, lema e unidade

Para entender o funcionamento do sistema de consultas por **elementos gramaticais** é esencial diferenciar entre os termos *palabra ortográfica*, *elemento gramatical*, *lema* e mais *unidade*.

A **palabra ortográfica** é a forma escrita entre dous espazos en branco. Empregando esta opción, o sistema non ten en conta abstraccións nin variacións de ningún tipo e así, se buscásemos por exemplo *biquei*, só obteríamos os resultados coincidentes con esa forma, pero non os casos de *biqueino*, *biqueite*, *biqueille* ou *biqueí*, se os houber.

O **lema** é o representante canónico dos elementos que se encadran baixo un mesmo paradigma e coincide, polo xeral, coa entrada do dicionario. Por exemplo, o lema *bicar* acolle toda a conxugación verbal do verbo *bicar*, de xeito que realizando as consultas por **lema** o sistema devolve todos os casos de todas as formas verbais que se engloban baixo ese paradigma verbal, independentemente de se a forma verbal concreta conforma unha palabra ortográfica illada (*biquei*, *bico*, *bicariamos...*) ou forma parte dun conglomerado constituído pola forma verbal e mais a segunda forma do artigo ou pronomes enclíticos (*bica-los*, *bícaa*, *bicábanse*, *bicabámonos* etc.). Este sería o campo que teríamos que cubrir co texto *ter* para, cunha soa consulta, obter todas as concorrencias do verbo *ter* coas que exemplificabamos máis arriba.

A **unidade** identifícase xeralmente con palabra ortográfica, menos naqueles casos en que a agrupación de máis dunha palabra dá lugar a unidades multipalabra, como son os diversos tipos de locución e, sobre todo, nomes propios e numerals constituídos por máis dunha palabra ortográfica. Así, ao lado das **unidades** *casa*, *quixemos*, *Xan*, *pouquiño*, *nunha*, *das*, *bícame* ou *tróuxenllelo*, constitúen tamén unha **unidade**, por exemplo, *Santiago de Compostela*, *San Cibrao das Viñas*, *a carón da* ou *vinte e sete mil trescentos corenta e tres*.

O **elemento gramatical** é a verdadeira unidade de análise do etiquetador, á que lle corresponde sempre unha etiqueta e un lema. A existencia das amálgamas, isto é, de dous ou máis constituíntes baixo unha mesma **palabra ortográfica** onde cada un dos compoñentes posúe unha etiqueta propia e un lema diferente forzan a súa aparición. A diferenza percíbese claramente cun exemplo: *deullo* é unha **unidade** composta polo **elemento gramatical** *deu*, cuxa etiqueta

correspondente 'Vei30s' indica que é a 3ª persoa do singular do pretérito de indicativo do lema *dar*, e mais polo **elemento gramatical lle** –etiquetado como pronome átono de terceira singular, masculino ou feminino en caso dativo (Rad3as), cuxo lema é *lle-* e, finalmente, polo **elemento gramatical o** –etiquetado como pronome átono de terceira singular masculino en caso acusativo (Raa3ms) cuxo lema é *o-*. É dicir, a palabra ortográfica *deullo*, non identificable cunha única clase de palabra nin asociada como un todo a ningunha etiqueta ou lema reais, constitúe en realidade tres elementos gramaticais con cadanseu par etiqueta e lema propios.

No caso das amálgamas coma *deullo*, palabra ortográfica e unidade coinciden; porén, a diverxencia é clara cando estamos ante un elemento gramatical multipalabra que forma parte dunha contracción. Por exemplo, *a pesar da* constitúe **unha unidade** conformada por **dous elementos gramaticais** distribuídos en **tres palabras ortográficas**:

unidade: *a pesar da*

elementos gramaticais: *a pesar de / a*

palabras ortográficas: *a / pesar / da*

Por si só, **unidade** non é un campo de consulta válido e non se permite realizar buscas cubrindo unicamente ese campo. A presenza do elemento **unidade** no sistema de consultas establécese para axudar a concretar máis as buscas e poder individualizar as ocorrencias dos **elementos gramaticais** que están amalgamados. Por exemplo, para comprobar en que casos se emprega o trazo coa denominada segunda forma do artigo⁵ basta con especificar a clase de palabra e o subtipo no campo **Etiqueta** (Dd*), cubrir o campo **Lema** co do artigo (*o*) e introducir en **Unidade** ***-l*** ou ***-***. Se nesta busca en **unidade** completásemos só con ***l***, o sistema devolvería todos os casos de segunda forma do artigo, con ou sen trazo, mais tamén as ocorrencias do artigo determinado que están implicadas en amálgamas con elementos gramaticais nos que figura un *l*, por exemplo en *relación coa*, *por culpa dos* ou *analiza-los*, pois cumpren cos criterios da busca que se está a realizar: todas as ocorrencias de artigo determinado en cuxa unidade haxa un *l*. Para evitar este tipo de ruído nos resultados, aconséllase realizar as consultas fundamentalmente por **elemento gramatical**, **etiqueta** e/ou **lema** e recorrer só á **unidade** para restrinxir as buscas.

A maiores, débese ter presente, tanto na realización das consultas como na observación dos resultados, que existen unidades con ambigüidade segmental que, en función do contexto no que se localicen, serán tratadas como unidades multipalabra ou como unidades consecutivas independentes. Por exemplo, se en elemento gramatical buscamos *pola*, o sistema vainos devolver só as concorrencias etiquetadas como substantivo, prescindindo dos casos da contracción da preposición e o artigo. Así mesmo, se procuramos o elemento gramatical *de acordo*, só nos aparecerán os casos de *de acordo* etiquetados como adverbio, pero non os de *de acordo con* caracterizados como unidade multipalabra prepositiva.

Non pode esquecerse tampouco que a etiquetaxe do corpus etiquetado automaticamente é, como o seu nome indica, totalmente automática, polo que é esperable que se produzan erros, ben na segmentación ben na etiquetaxe.

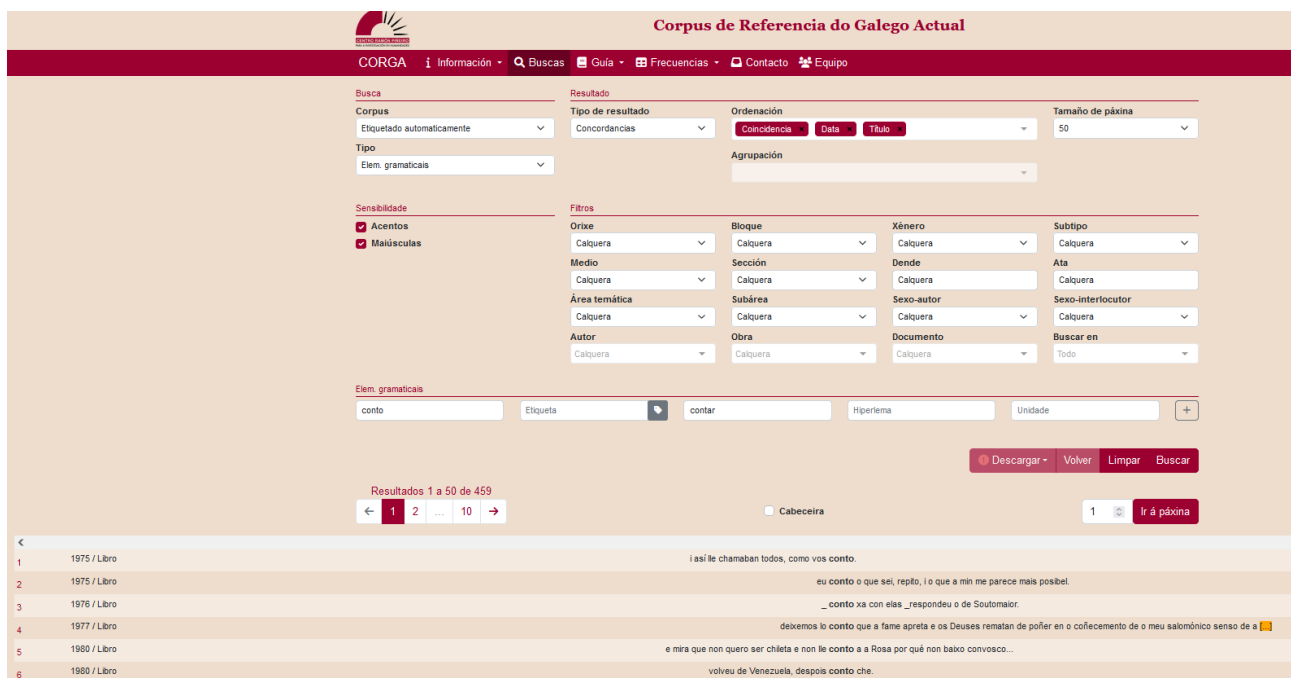
A posibilidade de empregar o lema facilita enormemente a recuperación de información no caso das clases de palabras que se caracterizan por presentar flexión (substantivos, adxectivos, verbos, artigos, posesivos...), pois neutraliza a variación debida ao incremento de desinencias, mais tamén anula a irregularidade nas raíces verbais entre o tema de presente e o de pretérito naqueles

5 Decidiuse deixar como forma do elemento gramatical a correspondente á segunda forma do artigo para facilitar a recuperación da información naqueles casos en que esta participaba, sabendo que a consulta por lema permite a súa agrupación. É por isto que no caso das contraccións nas que participa a segunda forma do artigo ou no conglomerado verbal coa segunda forma do pronome se mantén a grafía correspondente a esta: *lo, la, los, las*.

poucos verbos cuxo número está en relación inversamente proporcional ao seu uso: *ser, estar, ter, ir, ver, vir...*

A escolla de **Elementos gramaticais** nas opcións de busca activa os campos **elemento gramatical, etiqueta, lema, hiperlema e unidade**, co que se activa tamén a posibilidade de acceder ás características gramaticais. Así, se cubrimos en **elemento gramatical** *conto* e en **lema** especificamos *contar*, resolvemos o problema que formula a existencia do par homógrafo *conto* substantivo fronte a *conto* primeira persoa do singular do presente de indicativo do verbo *contar*, co que obtemos todas as concorrencias de *conto* verbo, independentemente da súa aparición en conglomerados de forma verbal con pronomes enclíticos. A pantalla co resultado das buscas a xeito de concordancias é semellante á que viamos anteriormente nas buscas por **Palabras ortográficas**: concordancia, ano e medio do documento no que se localiza a concorrencia e un díxito que indica o número do exemplo.

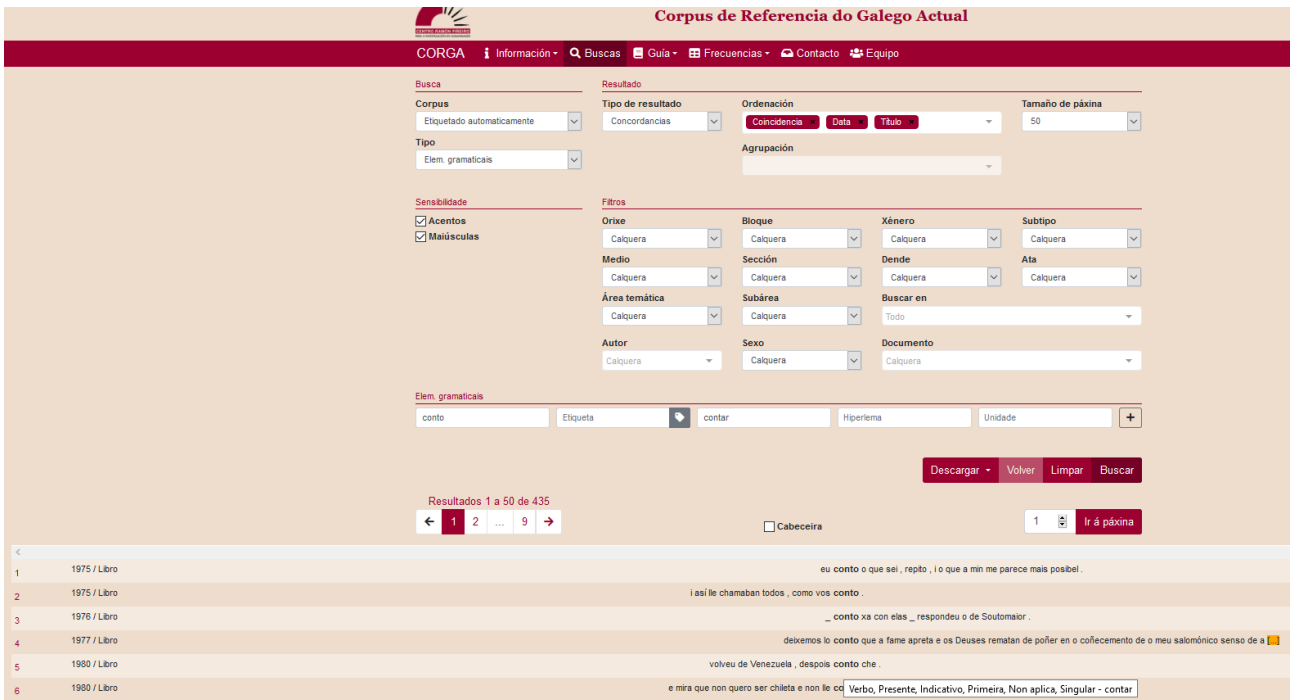
A diferenza estriba aquí, como se observa na figura seguinte, en que o texto da concordancia que se nos amosa co resultado obxecto da busca centrado e en negra está *tokenizado*, ou sexa, as amálgamas desagreganse nos seus elementos gramaticais constituíntes e as maiúsculas convencionais a comezo de enunciado convértense a minúsculas. Fixémonos na sexta concordancia, por exemplo, onde apreciamos que a consoante inicial de *volveu* está en minúsculas e logo vemos na parte final da secuencia algo en principio estraño para a ortografía convencional, pero totalmente lóxico na análise por elementos gramaticais: a desagregación de *cóntoche* en *conto* e mais *che*.



The screenshot shows the CORGA search interface. At the top, there's a navigation bar with 'CORGA' and various menu items. Below that, there are search filters for 'Busca', 'Tipo', 'Sensibilidade', and 'Filtros'. The search results are displayed in a table with columns for 'Resultado', 'Ordenación', and 'Tamaño de páxina'. The search term 'conto' is entered in the search box, and the results are sorted by 'Concordancia'. The search results are displayed in a list with columns for 'Ano', 'Documento', and 'Concordancia'. The search results are displayed in a list with columns for 'Ano', 'Documento', and 'Concordancia'. The search results are displayed in a list with columns for 'Ano', 'Documento', and 'Concordancia'.

Fig. 13. Concordancias coa desagregación en elementos gramaticais.

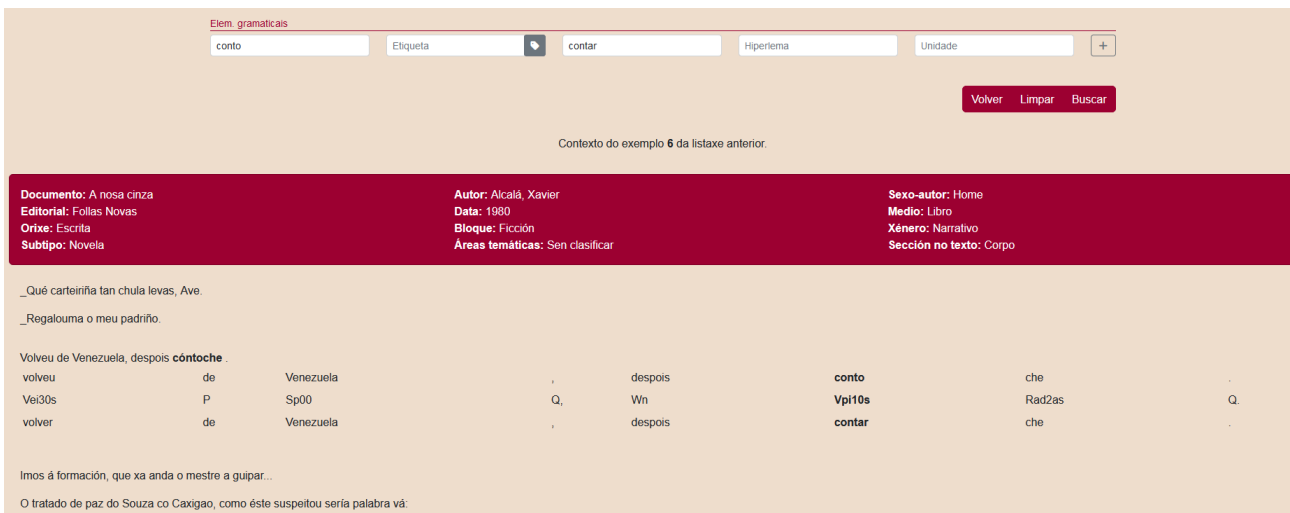
Se pasamos co punteiro do rato por riba do elemento obxecto da busca (*conto*) ou de calquera outro elemento presente nunha concordancia, emerxe un cadro de texto coa súa caracterización morfosintáctica. No caso que referimos *verbo, presente, indicativo, 1ª, xénero non aplica, singular*.



The screenshot shows the CORGA search interface. At the top, there is a navigation bar with 'CORGA' and links for 'Información', 'Buscas', 'Guía', 'Frecuencias', 'Contacto', and 'Equipo'. Below this, there are search filters for 'Busca' (Corpus, Tipo, Sensibilidade), 'Resultado' (Tipo de resultado, Ordenación, Tamaño de páxina), and 'Filtros' (Orixe, Bloque, Xénero, Subtipo, Medio, Sección, Dende, Área temática, Subárea, Autor, Sexo, Documento). A search bar contains 'conto' and 'contar'. Below the search bar, there are buttons for 'Descargar', 'Volver', 'Limpar', and 'Buscar'. The results section shows 'Resultados 1 a 50 de 435' and a list of search results with their respective contexts.

Fig. 14. Cadro emerxente coa caracterización morfosintáctica.

A outra diferenza salientable con respecto ás consultas por palabras ortográficas ten lugar cando prememos no número de orde dun exemplo concreto e accedemos á ampliación do seu contexto:



The screenshot shows the expanded context of a search result. At the top, there is a search bar with 'conto' and 'contar'. Below the search bar, there are buttons for 'Volver', 'Limpar', and 'Buscar'. The main content area shows the expanded context of the search result, including the document title, author, and other metadata. Below this, there is a table showing the context of the search result, with columns for the word, its frequency, and its context.

Documento: A nosa cinza	Autor: Alcalá, Xavier	Sexo-autor: Home
Editorial: Folhas Novas	Data: 1980	Medio: Libro
Orixe: Escrita	Bloque: Ficción	Xénero: Narrativo
Subtipo: Novela	Áreas temáticas: Sen clasificar	Sección no texto: Corpo

Contexto do exemplo 6 da listaxe anterior.

..._Qué carteiña tan chula levas, Ave.
...Regalouma o meu padriño.

Volveu de Venezuela, despois **contóche** .
volveu de Venezuela , despois conto che .
Vei30s P Sp00 Q, Wn Vpi10s Rad2as Q.
volver de Venezuela , despois contar che .

Imos á formación, que xa anda o mestre a guipar...
O tratado de paz do Souza co Caxigao, como éste sospeitou sería palabra vá:

Fig. 15. Contexto ampliado da concordancia 6 da imaxe anterior.

A información que contén cada liña, de arriba cara abaixo, é a seguinte:

- Cadro de texto cos metadatos pertinentes do documento correspondentes á concordancia concreta (documento, autor, sexo do autor, editorial, ano de publicación, medio, orixe, bloque, xénero, subtipo, área temática e sección no texto na que aparece).
- As dúas liñas seguintes recollen en cadansúa secuencia o texto que precede a oración na que aparece o elemento obxecto da consulta.

- A seguinte liña corresponde á secuencia na que aparece o elemento obxecto da consulta escrito en ortografía convencional. Repárese aquí na palabra destacada da imaxe: **cóntoche**.
- As tres liñas seguintes están interconectadas, dado que entre as tres identifican e caracterizan morfosintacticamente os elementos constitutivos do enunciado no que aparece o noso obxecto de consulta: a primeira delas recolle os elementos gramaticais; a do medio indica a etiqueta que corresponde ao elemento gramatical situado enriba dela e, por último, a terceira infórmanos do lema que corresponde a cada elemento gramatical. Na visualización apréciase con claridade a información relativa a cada elemento, malia apareceren en liñas distintas a etiqueta e o lema, debido á disposición gráfica escolleita, a xeito de columnas. Repárese en que a análise correspondente ao obxecto da busca aparece destacada en cada liña: **conto / Vpi10s / contar**. Por último, para ver aquí a que corresponde cada etiqueta basta situar o punteiro do rato enriba dunha para que emerxa o texto que a desenvolve.
- Pola súa banda, as dúas últimas liñas recollen en cadansúa secuencia o texto que segue a oración na que aparece o elemento obxecto da consulta.
- Finalmente, ao igual que viamos para as consultas por palabra ortográfica, no contexto incorpórase tamén a información relativa ao *interlocutor/falante* ao que se remite cada secuencia, se o documento no que se rexistra a ocorrencia é unha transcripción, unha obra de teatro, un guión ou unha entrevista. Inclúese, así mesmo, en idéntico formato ao de *interlocutor*, a información de se nun texto dramático a secuencia corresponde a unha *acoutación*. Ademais, todas as secuencias do contexto dunha ocorrencia que se localice nunha transcripción poden descargar ou reproducir o son dende alí premendo no punteiro da frecha que aparece en cada caso.

Sen dúbida, a opción de consulta máis rica é a que emprega a etiqueta morfosintáctica. Cada etiqueta desenvólvese clasificando inicialmente o elemento segundo a clase de palabra á que pertence, para deseguido, en función do tipo de palabra dado, atribuír os trazos morfosintácticos que caracterizan a clase cos valores concretos que corresponden segundo o contexto no que se insire o elemento ante o que esteamos.

Para evitar que o usuario teña que aprender o etiquetario (dispoñible en <http://corpus.cirp.gal/xiada/>), desenvolveuse un menú amigable que permite ir construíndo a etiqueta paso a paso. O menú reflicte os valores de todas as clases de palabras e dos atributos que aplican en cada unha delas e permite deixar en branco algúns dos trazos.

Para acceder a este menú de introdución da etiqueta hai que premer na tarxetiña sombreada situada xusto ao lado do campo **etiqueta**. A relación de valores comeza coa clase de palabra que se desexa e seguen despois as opcións existentes para as categorías gramaticais a medida que se van seleccionando os distintos trazos pertinentes para a clase de palabra escolleita.

Se continuamos co exemplo anterior, introducindo no campo **etiqueta** a clase de palabra *verbo*, como se amosa na figura seguinte, sen necesidade de especificar a caracterización morfosintáctica completa, desfariamos tamén a homografía existente entre o *conto* substantivo e o *conto* verbo, obtendo todos os casos nos que *conto* se adscribe á clase de palabras verbo:

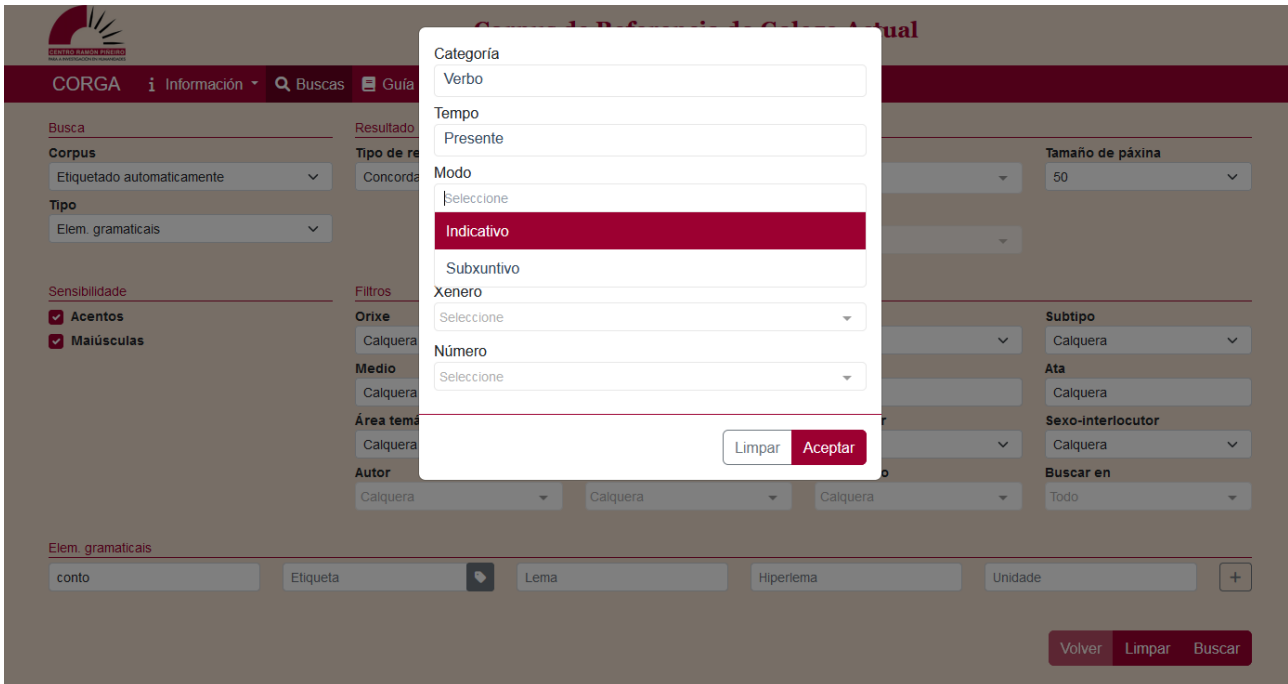


Fig. 16. Exemplo de definición de etiqueta a través do menú amigable.

Se a información relativa á etiqueta se introduce manualmente, esta visualizarase na caixiña correspondente:

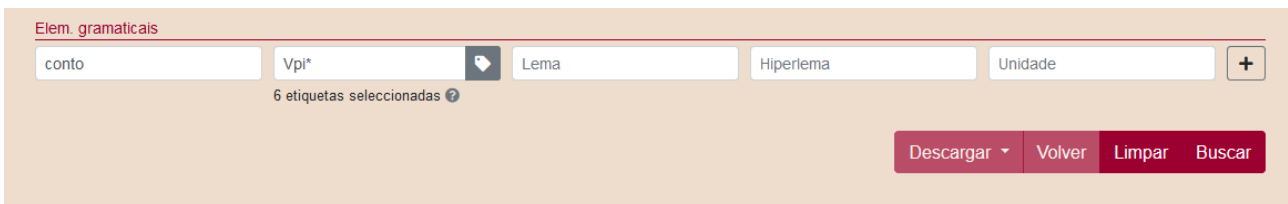


Fig. 17. Exemplo de visualización da etiqueta introducida manualmente.

En cambio, cando a información relativa á etiqueta se introduce mediante o menú amigable, esta non se reflicte na caixiña de texto, senón que para visualizala cómpre premer no ? situado debaixo e as etiquetas emerxerán nunha nova xanela:

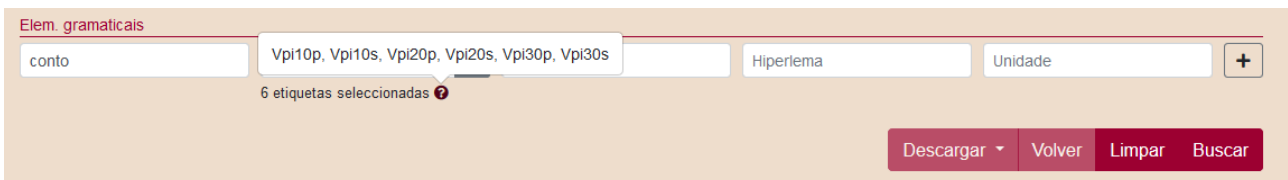


Fig. 18. Exemplo de visualización da etiqueta introducida mediante o menú amigable.

Con todo, o menú de inclusión de etiqueta non ten en conta o emprego dos operadores booleanos nin posibilita combinar estes con metacaracteres. Así, o menú de introdución fai posible, por exemplo, buscar calquera das formas verbais de infinitivo conxugado, só as formas plurais, ou as formas singulares e plurais correspondentes a unha persoa concreta, poñamos por caso, mais non permite excluír destas unha etiqueta concreta –a relativa ao infinitivo non persoal, V0f000⁶– ou

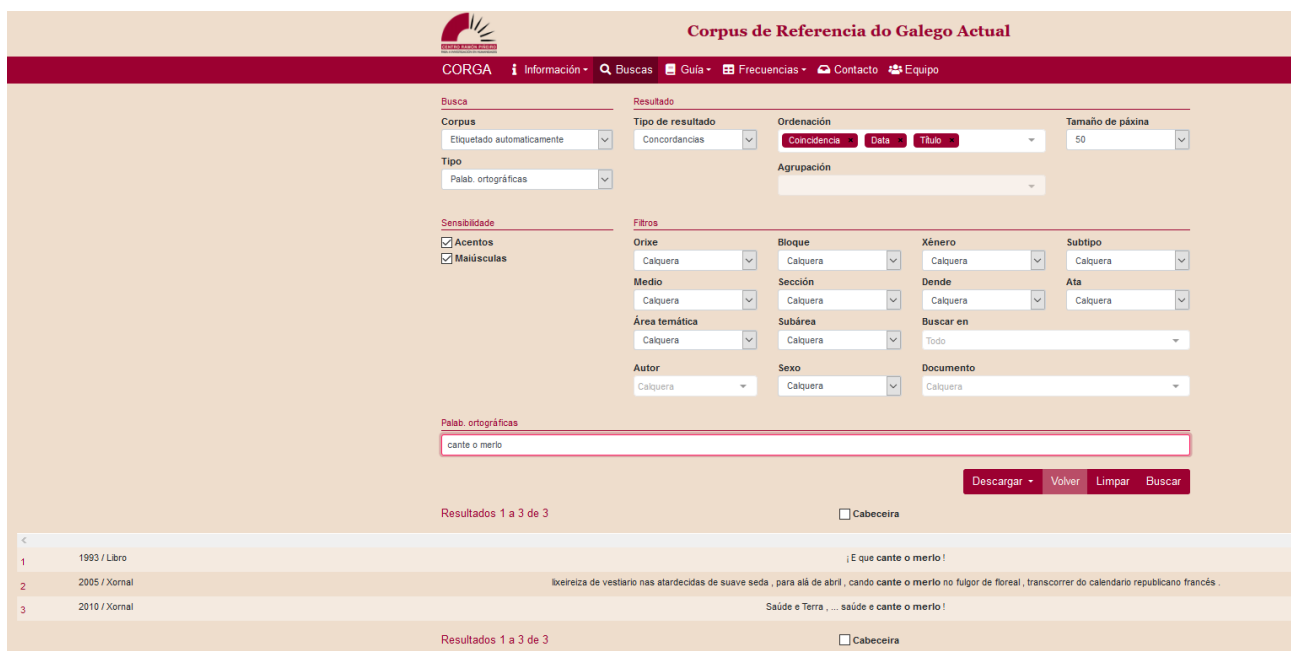
6 Dada a confluencia formal entre as formas do infinitivo non persoal e da primeira e terceira persoas do infinitivo flexionado, decidiuse non desambigualas, de maneira que todas elas se recuperarán baixo a etiqueta V0f000.

agrupar etiquetas nas que estean implicados varios trazos gramaticais –só as formas da 2ª e 3ª persoas plurais–. Non poderemos, polo tanto, recuperar cunha única consulta todas as formas verbais de infinitivo conxugado mediante o menú de inclusión. Para poder realizar consultas desas características, o campo da etiqueta debe formalizarse de xeito manual segundo as instrucións recollidas no apartado [5.1. Metacaracteres e operadores booleanos](#).

A posibilidade de acceder aos datos das concorrencias a través de etiquetas, podendo prescindir dos lemas aos que pertencen e dos elementos gramaticais nos que se concretan as etiquetas é, sen lugar a dúbida, a parte máis abstracta da ferramenta, pero tamén a máis potente, pois permite a obtención de datos operando con abstraccións gramaticais. Por exemplo, pódense tirar todos os casos contidos no corpus do infinitivo conxugado e estudar despois os datos véndoos en **concordancias**; e tamén se pode ir máis alá e extraer só a segunda persoa plural do infinitivo conxugado. A enorme riqueza deste sistema é que permite que sexa o usuario quen defina a complexidade da procura, podendo limitarse ás clases de palabras ou pola contra interesarse por algún dos valores concretos dos trazos que son pertinentes para a caracterización morfosintáctica dun determinado tipo de palabra. Ou sexa, pódense recuperar as ocorrencias de todos os substantivos femininos plurais, prescindindo de se son comúns ou propios (*S?fp* no campo **etiqueta**), ou ben pedir todos os casos de femininos plurais (**fp* no campo **etiqueta**, o que incluíría casos de todas as clases de palabras nas que estes trazos son pertinentes: substantivo, adxectivo, verbo, artigo, demostrativo...).

5.4. Consulta por elementos sucesivos

O sistema permite as buscas de elementos sucesivos (ata 5), tanto nas consultas por **Palabras ortográficas** como por **Elementos gramaticais**. No primeiro caso introdúcese o segmento de consulta no campo **Texto**, por exemplo a fórmula de despedida *cante o merlo*, a cal se documenta 2 veces:



The screenshot shows the CORGA web interface. At the top, there is a navigation bar with 'CORGA' and links for 'Información', 'Buscas', 'Guía', 'Frecuencias', 'Contacto', and 'Equipo'. Below this, there are search filters for 'Busca' (labeled 'Resultado'), 'Tipo de resultado' (set to 'Concordancias'), 'Ordenación' (set to 'Coincidencia'), and 'Tamaño de páxina' (set to '50'). There are also dropdowns for 'Etiquetado automaticamente' and 'Tipo' (set to 'Palab. ortográficas').

Under 'Sensibilidade', there are checkboxes for 'Acentos' and 'Maiúsculas', both checked. The 'Filtros' section includes: 'Orixe' (Calquera), 'Bloque' (Calquera), 'Xénero' (Calquera), 'Subtipo' (Calquera), 'Medio' (Calquera), 'Sección' (Calquera), 'Dende' (Calquera), 'Ata' (Calquera), 'Área temática' (Calquera), 'Subárea' (Calquera), 'Buscar en' (Todo), 'Autor' (Calquera), 'Sexo' (Calquera), and 'Documento' (Calquera).

The search field 'Palab. ortográficas' contains the text 'cante o merlo'. Below the search field are buttons for 'Descargar', 'Volver', 'Limpar', and 'Buscar'. The results section shows 'Resultados 1 a 3 de 3' and a 'Cabeceira' checkbox. The results table has 3 rows:

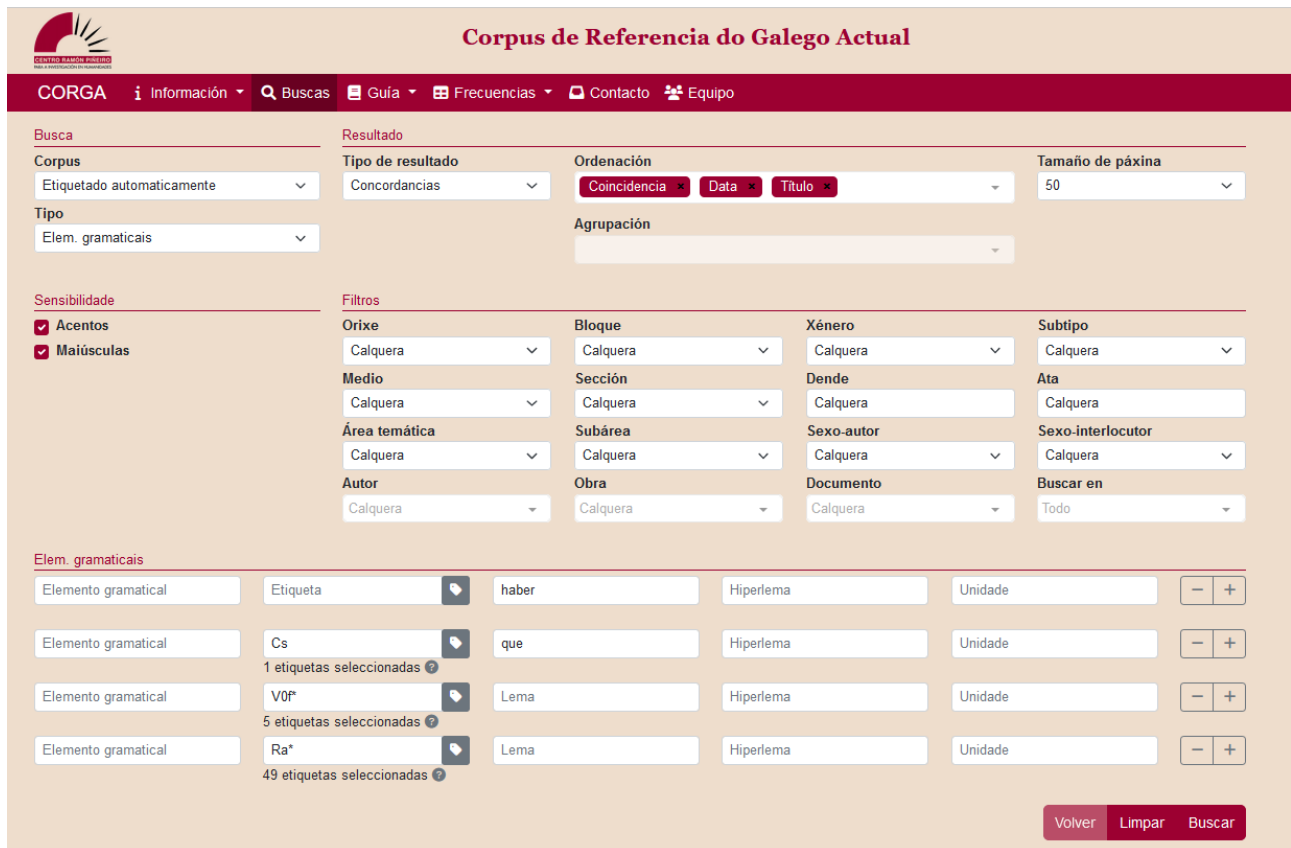
1	1993 / Libro	E que cante o merlo !
2	2005 / Xornal	breixeira de vestiario nas atardecidas de suave seda , para alá de abril , cando cante o merlo no fulgor de forestal , transcorrer do calendario republicano francés .
3	2010 / Xornal	Saúde e Terra , ... saúde e cante o merlo !

At the bottom, it shows 'Resultados 1 a 3 de 3' and another 'Cabeceira' checkbox.

Fig. 19. Consulta de elementos sucesivos por palabras ortográficas.

En tanto que na segunda modalidade, cada vez que queiramos introducir características correspondentes a un novo elemento, temos que premer no + que está xusto despois do campo **Unidade**, enriba do botón de **Buscar**. Non é preciso que completemos a información correspondente

aos cinco campos que aparecen para cada elemento: **elemento gramatical**, **etiqueta**, **lema**, **hiperlema** e **unidade**, podendo combinalos como se queira. Así, para comprobar cal é a posición que ocupa preferentemente o pronome átono na perífrase de obrigatoriedade *haber + que + infinitivo* cóbrease na liña do primeiro elemento o lema *haber*, na liña do segundo elemento inclúese a etiqueta Cs (conxunción subordinante) e mais o lema *que*; na terceira liña especificase a etiqueta de infinitivo (*VO**), e finalmente, na cuarta liña precisase a etiqueta de pronome átono (*Ra**).



The screenshot shows the 'Corpus de Referencia do Galego Actual' search interface. The search criteria are: Corpus: 'Etiquetado automaticamente', Tipo: 'Elem. gramaticais', Resultado: 'Concordancias', Ordenación: 'Coincidencia', and Tamaño de páxina: '50'. The filters section includes 'Sensibilidade' (checked for 'Acentos' and 'Maiúsculas'), 'Filtros' (Orixe, Medio, Área temática, Autor, Bloque, Sección, Subárea, Obra, Xénero, Dende, Sexo-autor, Documento, Subtipo, Ata, Sexo-interlocutor, Buscar en), and 'Elem. gramaticais' (Elemento gramatical, Etiqueta, Lema, Hiperlema, Unidade). The search results table shows four rows of results for the query 'haber + que + infinitivo'.

Elemento gramatical	Etiqueta	Lema	Hiperlema	Unidade
haber		haber	Hiperlema	Unidade
que	Cs	que	Hiperlema	Unidade
Lema	VO*	Lema	Hiperlema	Unidade
Lema	Ra*	Lema	Hiperlema	Unidade

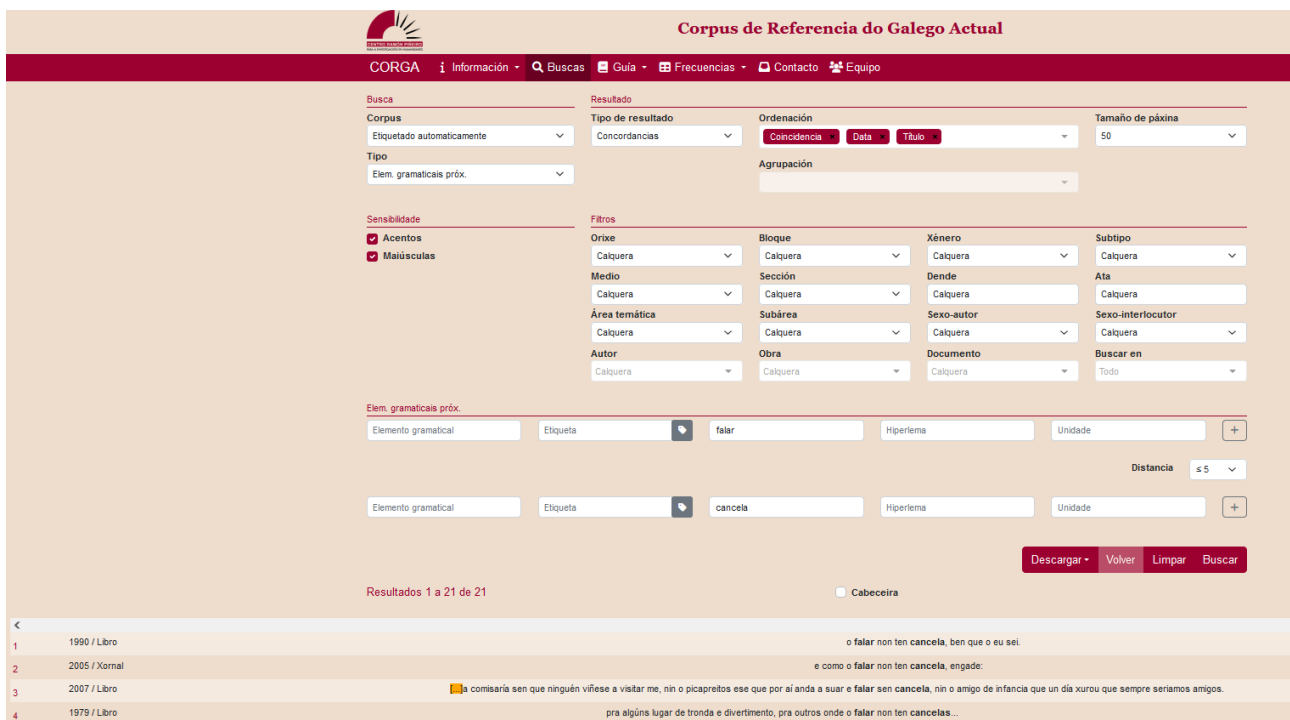
Fig. 20. Consulta por elementos gramaticais sucesivos.

Basta con repetir a busca mudando a posición da cuarta liña, a relativa ao pronome átono, para a segunda e terceira liñas sucesivamente ata esgotar as posibles posicións. Por suposto, tamén podemos pescudar cal é o tempo verbal máis empregado nesta perífrase no corpus detallando na liña do verbo auxiliar a etiqueta pertencente ao presente, futuro, pretérito, copretérito, pospretérito... Introducindo na etiqueta *Vpi** (clase verbo, tempo presente, modo indicativo) obteríamos un total de *x* casos, co que se temos en conta que a frecuencia desta perífrase de obrigatoriedade é de *x* casos por millón, as concorrencias en terceira persoa do singular do presente de indicativo supoñerían unha frecuencia de *x*. Tamén poderíamos obter o dato inverso: cantos casos hai da perífrase que non teñen como verbo auxiliar a terceira de singular do presente de indicativo, só con especificar no campo do **elemento gramatical** *!hai*, ou *sexa*, calquera das formas verbais de *haber* que poden aparecer nesa construción menos *hai*.

5.5. Consulta por proximidade

As procuras anteriores exemplifican o emprego das expresións regulares para substituír caracteres e mais para restrinxir a obtención de resultados cunha determinada forma verbal. O seu emprego en combinación con etiquetas e elementos gramaticais dá boa mostra da potencia das

buscas. Non obstante, presentan unha pexa importante, e é que a consulta por elementos gramaticais sucesivos esixe que se respecte o padrón numérico da estrutura que estamos buscando. Seguindo co exemplo anterior, se algunha ocorrencia presentase máis dun pronome enclítico non se tería en conta nos resultados, porque só pediamos un pronome enclítico, e ocorrencias coma *habirá que encargárllelas, hai que darlla, hai que reconecerllo...* non imos obtelas coa consulta realizada. Para paliar esa pexa xorden as consultas por proximidade, de xeito que dándolle ao sistema **dúas palabras ortográficas** ou **dous elementos gramaticais** poidamos recuperar todos os casos nos que eses dous elementos se atopen a unha distancia X (o usuario decide, cun valor de entre 1 e 10, a distancia, exacta ou non, que se debe ter en conta para extraer as coaparicións). Así, se no primeiro elemento gramatical cubrimos en lema con *falar*, no segundo completamos tamén en lema con *cancela* e especificamos que a distancia en que deben aparecer ambos os dous lemas é igual ou inferior a 5 (≤ 5), imos obter os casos *o falar non ten cancelas, o falar non ten cancela, o falar non che ten cancelas, falar sen cancela, falar e prometer non ten cancelas*:



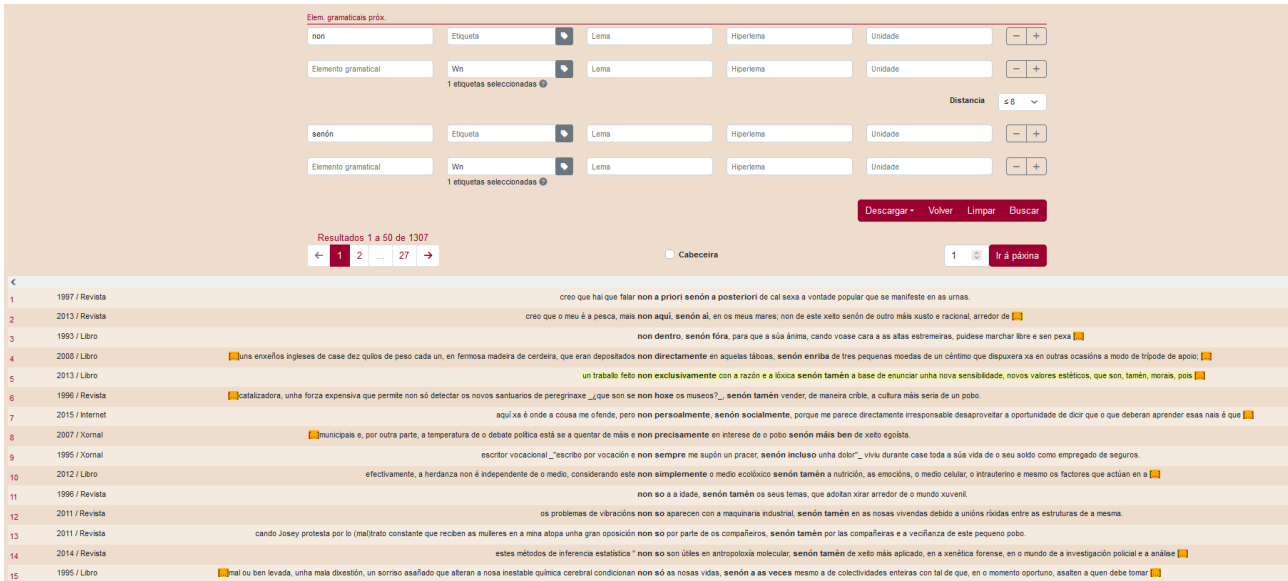
The screenshot shows the CORGA search interface. At the top, there's a navigation bar with 'CORGA', 'Información', 'Buscas', 'Guía', 'Frecuencias', 'Contacto', and 'Equipo'. Below this, there are search filters for 'Busca' (Corpus, Tipo), 'Resultado' (Tipo de resultado, Ordenación, Agrupación), and 'Sensibilidade' (Acentos, Maiúsculas). A grid of filters for grammatical categories is visible, including Orixe, Bloque, Xénero, Subtipo, Medio, Sección, Dende, Ata, Área temática, Subárea, Sexo-autor, Sexo-interlocutor, Autor, Obra, Documento, and Buscar en. The search input area shows two queries: 'falar' and 'cancela', both with a distance of ≤ 5 . Below the search area, there are buttons for 'Descargar', 'Volver', 'Limpar', and 'Buscar'. The results section shows 'Resultados 1 a 21 de 21' and a 'Cabeceira' checkbox. The first result is '1990 / Libro' with the text 'o falar non ten cancela, ben que o eu sei.'

Fig. 21. Procura por elementos gramaticais próximos.

Nas buscas por proximidade hai que ter en conta, por unha banda, que se optamos por un dígito sen que estea precedido do símbolo ' \leq ', estamos pedindo que se nos devolvan todos os casos nos que os dous elementos próximos en cuestión estean á distancia exacta sinalada, de xeito que na busca anterior se en vez de ' ≤ 5 ' optásemos por '5', só nos devolvería *falar e prometer non ten cancelas*. Pola outra banda, débese ter presente que canto maior sexa a distancia establecida entre as dúas palabras ou elementos gramaticais, a non ser que sexa unha estrutura de uso moi reducido, maiores posibilidades hai de que entre os resultados apareza ruído, ou sexa, casos coincidentes formalmente co obxecto da procura, pero non co que se pretendía buscar.

A partir da versión 4.0 o número de palabras ou elementos gramaticais que se sitúan en cada un dos extremos amplíase de 1 ata 5, incrementando así as posibilidades de recuperación de información e facendo máis flexible a aplicación. Esta modalidade facilitaríanos a obtención, por exemplo, das estruturas correctivas constituídas polo modelo *non só... senón tamén*, ou mesmo de

non seguido dun adverbio nuclear situado a unha distancia de ata 8 elementos de *senón* seguido pola súa vez doutro adverbio nuclear. Esta última procura é a reflectida na figura que segue:



The screenshot shows the search interface for 'Elementos gramaticais próximos'. The search terms are 'non' and 'senón'. Filters include 'Etiqueta', 'Lema', 'Hiperlema', and 'Unidade'. The distance is set to '≤ 8'. The results list 15 items, each with a year and source, and a text excerpt containing the search terms.

Fig. 22. Procura por elementos gramaticais próximos.

Tamén nesta modalidade de consulta é posible introducir os metacaracteres e as expresións regulares para ampliar ou restrinxir certos resultados na recuperación de información. Pensemos na enorme utilidade de poder introducir, por exemplo, os lemas *máis|menos* e *ca|que* respectivamente en cada un dos extremos de proximidade se a estrutura que desexamos estudar son as comparativas de superioridade e inferioridade.

5.6. O hiperlema

Dada a enorme variación ortográfica que caracteriza, maiormente pero non só, os textos anteriores a 1982 e mais os cambios na representación gráfica dalgunhas unidades provocados por modificacións na normativa oficial, así como a existencia no marco desta última de duplicidades gráficas (entre elas pódense salientar os sufixos *-ble* e *-bel* ou *-aría* e *-ería*), na modalidade de consultas por **Elementos gramaticais** e **Elementos gramaticais próximos** a partir da versión 3.1 agrégase un parámetro máis para minimizar a variación formal entre lemas, se a quen realiza a procura lle prouguer: o **hiperlema**.

Así, se se quere pescudar que verbos se constrúen cun complemento coa preposición *ata*, cómpre ter en conta que esta coexiste nos textos coas variantes gráficas *até*, *ate*, *hasta* e *hastra*. Cada unha delas remítese na base de datos léxica ou lexicón a un lema coincidente coa forma, o que facilita a súa recuperación individualizada e, en consecuencia, a obtención de datos e frecuencias por cada unha desas grafías. Agora ben, o feito de crear un hiperlema *ata* que agrupa os lemas *ata*, *até*, *ate*, *hasta* e *hastra* permite neutralizar a variación ortográfica existente entre eles e facilita a obtención dos datos cando o verdadeiramente importante é o padrón estrutural e non o lema concreto.

Pensemos agora en *ditar*, hiperlema que en XIADA acolle os lemas *ditar* e *dictar*. Supoñamos que remitiramos as dúas conxugacións, a de *ditar* e a de *dictar* ao lema *ditar*, o único recollido nos dicionarios actuais. Como recuperaríamos só as ocorrencias de *ditar* e non as de *dictar*? Habería que buscar cada unha das formas verbais individualizadas (*dito*, *ditas*, *ditaba*, *ditei*

etc.) e ter en conta que poderían aparecer con pronomes enclíticos, o que obrigaría a buscar cada unha das combinacións posibles (*ditouno, diteiche, diteicho* etc.) e convertería a procura nun labor hercúleo.

Dende un punto de vista nocional, é preciso distinguir así mesmo entre lema e hiperlema. *Grosso modo*, o lema é a forma base sen os morfemas gramaticais. Así, tradicionalmente, as formas verbais agrúpanse baixo o infinitivo ou os adxectivos con oposición xenérica e numérica recóllense baixo a súa forma masculina singular. Pola súa banda, o hiperlema é un concepto relativamente novo, xerarquicamente superior, que nace para agrupar lemas que, no noso caso, se caracterizan por manifestaren entre eles unha relación de semellanza formal con pequenas variacións gráficas. Variacións no mesmo sufixo (*cafetería-cafetería; amable-amábel*), cambio de conxugación (*combaten-combatir, discorrer-discurrir*), presenza de grupo culto (*dictar, construcción, rector*) ou supresión (*ditar, construción*) e vocalización deste (*reitor*), ou vacilacións no vocalismo (*adobo-adubo, entroido-antroido*) exemplifican algunhas das variacións ortográficas que temos en conta no establecemento do hiperlema, coincidente maioritariamente coa forma á que a normativa oficial lle concede primacía, ben explícita ou ben porque a usa nos seus textos ou é a que aparece definida no DRAG. Porén, no lexicón hai lemas que non están avalados polas autoridades académicas, mais cuxa existencia xustifica o seu emprego nos textos recollidos no CORGA. Así, nin a *plantear*, *plantexar* ou *prantexar* os avala ningunha autoridade, mais nos documentos concorren formas conxugadas destes lemas, polo que é forzosa a súa introdución no lemario co fin de facilitar a recuperación de información, e con esta mesma finalidade elevamos á categoría de hiperlema *plantexar*.

Non se unifican mediante o hiperlema, non obstante, sufixos diferentes (*alargo e alargamento; canadense e canadiano*), formas patrimoniais e formas pertencentes a outras linguas (*coxín de seguridade vs. airbag; xeonllo, xionllo vs. rodilla*) ou sinónimos sen relación de semellanza gráfica (*croucheira, nogueira*). Naturalmente, o hiperlema aplícase tamén no caso das unidades multipalabra, de xeito que a través do hiperlema *grazas a* recuperaremos non só todas as ocorrencias da locución prepositiva *grazas a* senón tamén as de *gracias a*.

Por último, cómpre aínda outra puntualización sobre o establecemento do hiperlema, agora en relación coa homonimia, e é que no lexicón non recollemos información semántica e en consecuencia non distinguimos entre lemas homónimos, polo que, se un deles deba remitirse a un hiperlema diferente, non vai facerse pola confluencia co homónimo. Percíbese mellor cun exemplo. Nos textos do CORGA conflúen o lema non normativo *gabela* (na súa significación de porción de herba, palla etc.) e mais o lema *gabela* ‘imposto’, que conforman no lexicón un único lema, dado que ambos os dous son substantivos femininos e posúen os mesmos elementos paradigmáticos con idénticas caracterizacións morfolóxicas: *gabela* é substantivo común feminino singular e *gabelas*, mesma definición só que plural, independentemente de que correspondan ao ‘imposto’ ou ao ‘monllo’. Pois ben, esta confluencia, e a non inclusión de acepcións ou outro tipo de información semántica no lexicón, provoca que non remitamos o *gabela* de porción de herba ou monllo ao hiperlema *gavela*, pois iso implicaría que todos os *gabela* (porción de herba e tributo) se agrupasen baixo *gavela*, o que sería falso, xa que o tributo necesariamente ten que remitirse á forma con ‘b’.

Antes de entrar en como facer uso do parámetro **hiperlema** na aplicación de consulta, é preciso subliñar que o funcionamento dos metacaracteres e das expresións regulares mantense inalterado, de xeito que se estende, nas mesmas condicións que describimos máis arriba no apartado [Metacaracteres e operadores booleanos](#), ao hiperlema. O mesmo ocorre co resto de parámetros e condicionantes, ben sexa a sensibilidade a acentos gráficos e maiúsculas, as posibles restricións por algún dos criterios clasificatorios dos documentos ou a selección dun subcorpus virtual atendendo a unha parte estrutural concreta dos documentos ou a áreas temáticas específicas. Nada se modifica

nese sentido na aplicación.

Respecto de procuras nas que interveña o hiperlema, guíanos o propio sistema, que ofrece un campo específico para completalo, como se amosa na figura seguinte:

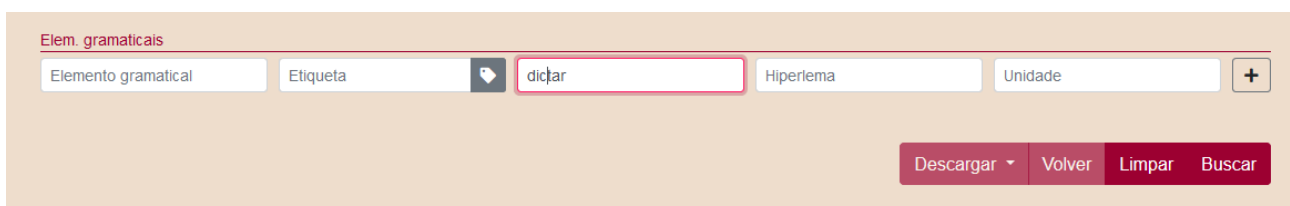


The screenshot shows the 'Corpus de Referencia do Galego Actual' search interface. It features a navigation bar with 'CORGA', 'Información', 'Buscas', 'Guía', 'Frecuencias', 'Contacto', and 'Equipo'. The main search area is divided into several sections: 'Busca' (search criteria), 'Resultado' (result options), 'Sensibilidade' (sensitivity), and 'Filtros' (filters). The 'Busca' section includes 'Corpus' (Etiquetado automaticamente), 'Tipo' (Elem. gramaticais), and 'Sensibilidade' (Acentos, Maiúsculas). The 'Resultado' section includes 'Tipo de resultado' (Concordancias), 'Ordenación' (Coincidencia, Data, Título), and 'Tamaño de páxina' (50). The 'Filtros' section includes 'Orixe', 'Bloque', 'Xénero', 'Subtipo', 'Medio', 'Sección', 'Dende', 'Ata', 'Área temática', 'Subárea', 'Sexo-autor', 'Sexo-interlocutor', 'Autor', 'Obra', 'Documento', and 'Buscar en'. The search input area at the bottom includes 'Elemento gramatical', 'Etiqueta', 'Lema', 'Hiperlema', and 'Unidade', with a search button and 'Volver', 'Limpar', and 'Buscar' buttons.

Fig. 23. Procura por elementos gramaticais.

As procuras nas modalidades de **Elementos gramaticais** e **Elementos gramaticais próximos** ofrecen as seguintes opcións para completar a caixa do **lema** e **hiperlema**:

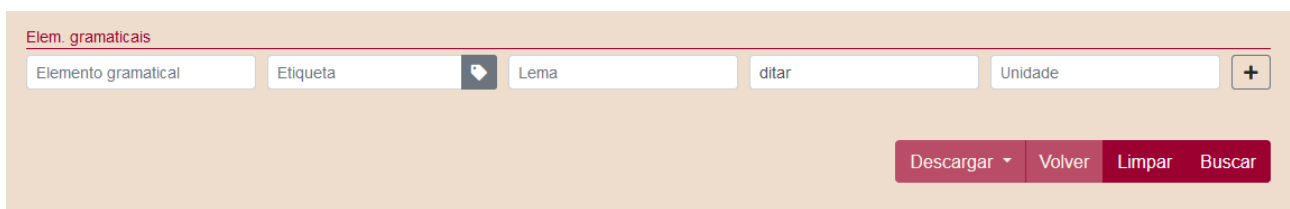
1) Buscar só polo lema: *dictar*. Devolve todas as formas verbais de *dictar*, constituídas por unha unidade gráfica ou formando parte dun conglomerado de forma verbal e pronome enclítico e/ou segunda forma do artigo:



The screenshot shows the search interface with the 'Elem. gramaticais' section selected. The search input area has 'Lema' set to 'dictar'. The search button is highlighted, and the 'Buscar' button is visible.

Fig. 24. Procura polo lema *dictar*.

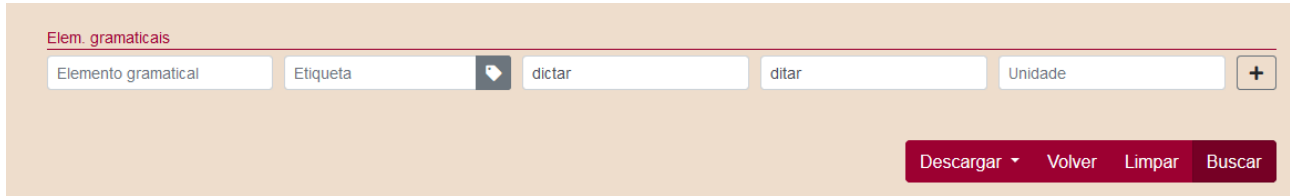
2) Buscar polo hiperlema: *ditar*. Devolve todas as formas verbais de *dictar* e mais as de *ditar*, constituídas por unha unidade gráfica ou formando parte dun conglomerado de forma verbal e pronome enclítico e/ou segunda forma do artigo:



The screenshot shows the search interface with the 'Elem. gramaticais' section selected. The search input area has 'Lema' set to 'ditar'. The search button is highlighted, and the 'Buscar' button is visible.

Fig. 25. Procura polo hiperlema *ditar*.

3) Buscar polo lema *dictar* e o hiperlema *ditar*. Compórtase coma en 1) e devolve só as formas da conxugación de *dictar*.

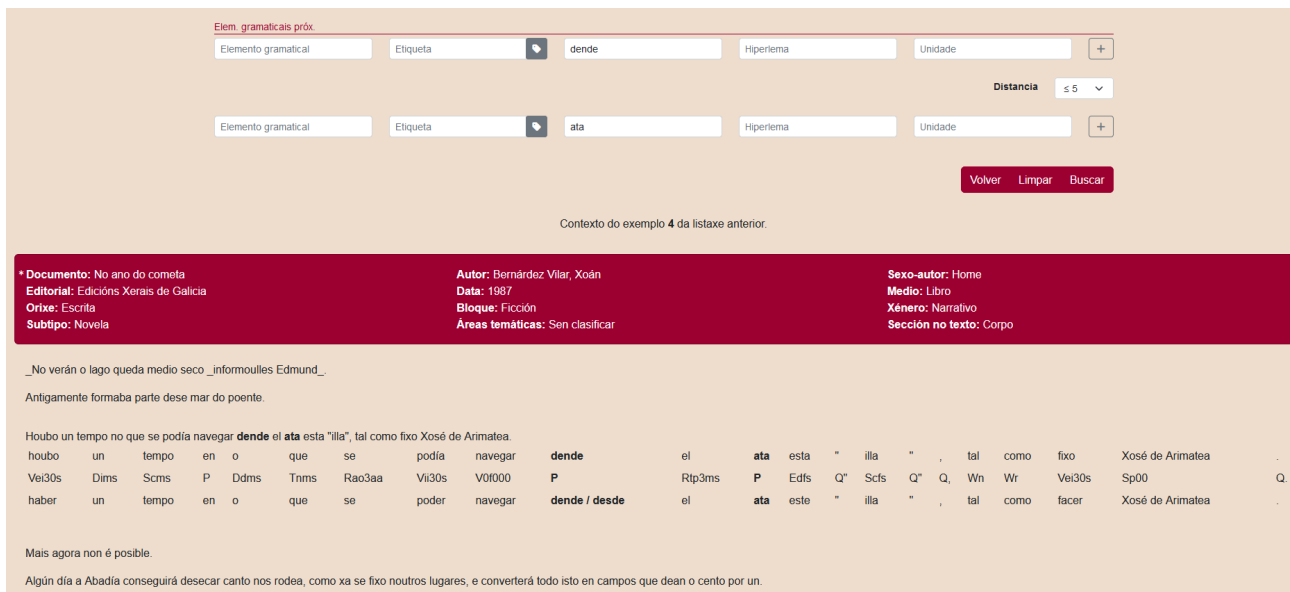


The screenshot shows a search interface with the following elements:

- Header: Elem. gramaticais
- Form fields: Elemento gramatical, Etiqueta, dictar, ditar, Unidade.
- Buttons: Descargar, Volver, Limpar, Buscar.

Fig. 26. Procura polo lema *dictar* e o hiperlema *ditar*.

O hiperlema especificase unicamente se difire do lema e o formato escollido para representalo na análise completa, á que se accede dende o contexto, é lema/hiperlema, como pode observarse na análise de *Houbo un tempo no que se podía navegar dende el ata esta "illa", tal como fixo Xosé de Arimatea*. que recollemos na seguinte imaxe. Nela, apréciase na terceira liña para o elemento *dende*, clasificado na segunda como P (Preposición), a información *dende/desde*, que remite o elemento ao lema *dende* e ao hiperlema *desde*, mais non repite o hiperlema en *ata* por seren coincidentes.



The screenshot shows the search interface with the following elements:

- Header: Elem. gramaticais próx
- Form fields: Elemento gramatical, Etiqueta, dende, Hiperlema, Unidade.
- Buttons: Volver, Limpar, Buscar.
- Contexto do exemplo 4 da listaxe anterior.
- Document information:
 - * Documento: No ano do cometa
 - Editorial: Edicións Xerais de Galicia
 - Orixe: Escrita
 - Subtipo: Novela
 - Autor: Bernárdez Vilar, Xoán
 - Data: 1987
 - Bloque: Ficción
 - Áreas temáticas: Sen clasificar
 - Sexo-autor: Home
 - Medio: Libro
 - Xénero: Narrativo
 - Sección no texto: Corpo
- Text:

_No verán o lago queda medio seco _informoulles Edmund_
Antigamente formaba parte dese mar do poente.

Houbo un tempo no que se podía navegar **dende** el **ata** esta "illa", tal como fixo Xosé de Arimatea.

houbo	un	tempo	en	o	que	se	podía	navegar	dende	el	ata	esta	"	illa	"	,	tal	como	fixo	Xosé de Arimatea	.
Vei30s	Dims	Scms	P	Ddms	Tnms	Rao3aa	Vii30s	V0i000	P	Rtp3ms	P	Edfs	Q"	Scfs	Q"	Q	Wn	Wr	Vei30s	Sp00	Q.
haber	un	tempo	en	o	que	se	poder	navegar	dende / desde	el	ata	este	"	illa	"	,	tal	como	facer	Xosé de Arimatea	.

Mais agora non é posible.

Algún día a Abadía conseguirá desecar canto nos rodea, como xa se fixo noutros lugares, e converterá todo isto en campos que dean o cento por un.

Fig. 27. Mostra da visualización dun hiperlema.

Débase ter presente que, en xeral, as consultas por lema ou hiperlema presentan unha limitación: se no lexicón non consta o lema, o etiquetador, a través do módulo de adiviñación, vai aventurar unha etiqueta tendo en conta a terminación da propia palabra descoñecida e das análises que constitúen o seu contexto, mais non ofrecerá un lema, a menos que fose obxecto de lematización automática, como veremos [no apartado 5.6.1](#). Isto tradúcese na aplicación de consulta do seguinte xeito:

1) Se se realiza unha consulta por un lema ou hiperlema que non está no dicionario/lexicón de XIADA, o sistema indicará que non atopa resultados, inda que o corpus rexistre ocorrencias dalgunha forma do paradigma dese lema ou hiperlema.

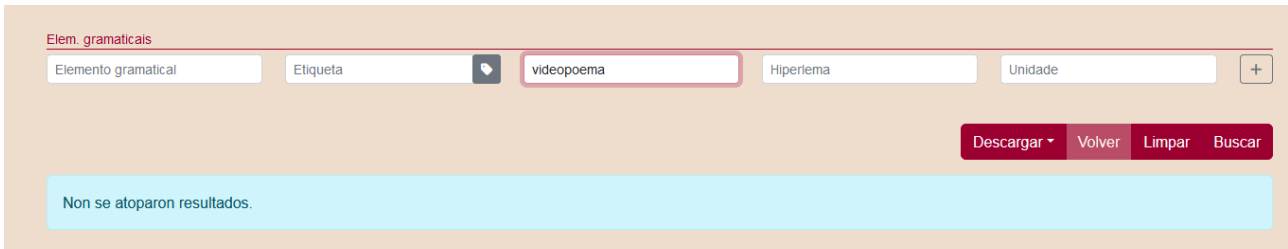


Fig. 28. Exemplo de procura dun lema.

2) Se se realiza a consulta por etiqueta ou elemento gramatical e se accede á análise completa da secuencia a través do contexto, na terceira liña, a correspondente aos lemas, aparecerá para os elementos descoñecidos un asterisco:

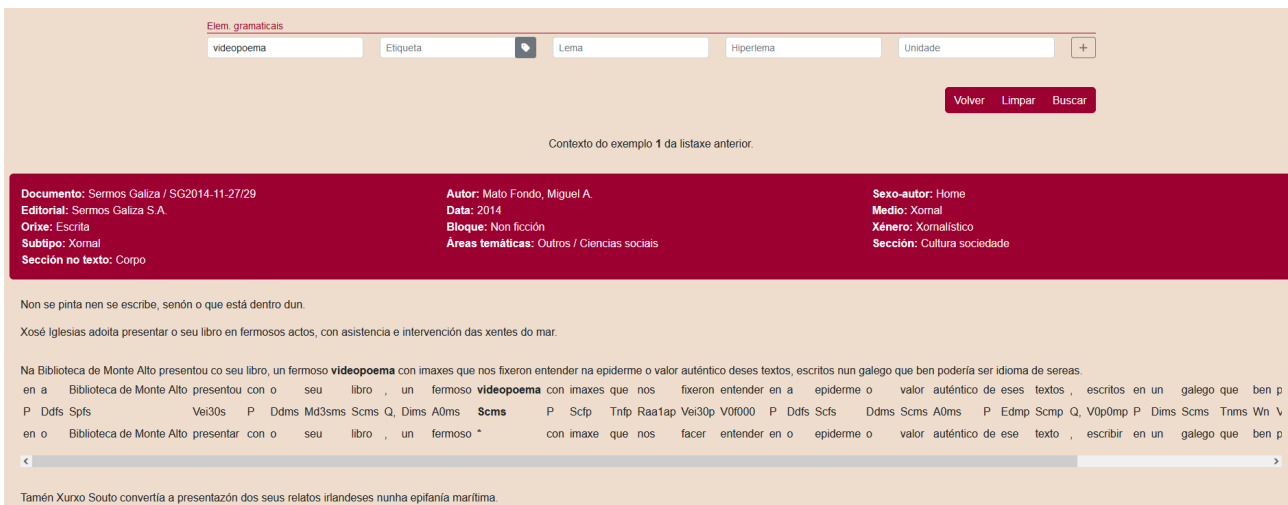


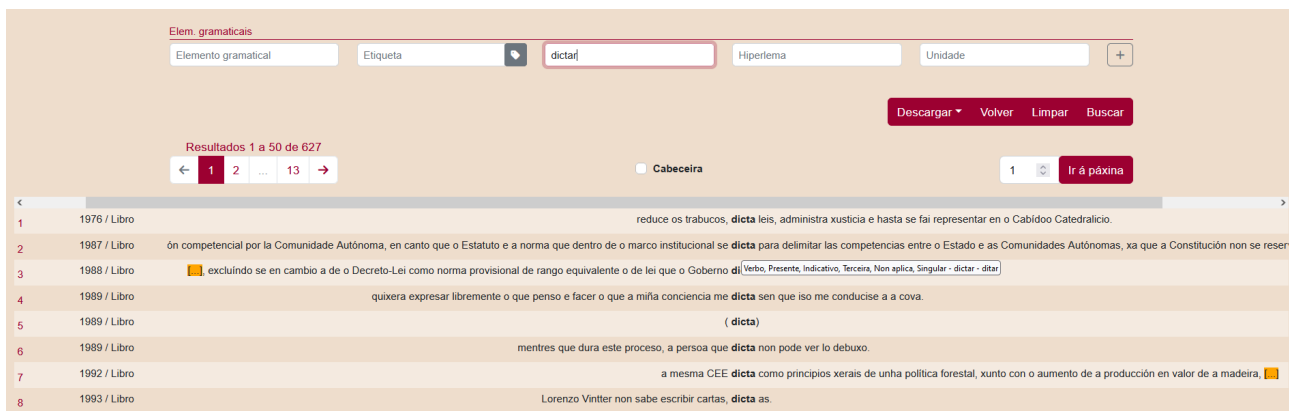
Fig. 29. Exemplo de elemento analizado que non presenta lema.

Lembremos que o hiperlema se especifica na análise unicamente se difire do lema, como pode observarse na figura 30, análise visible dende o contexto. Ou sexa, cando lema e hiperlema converxen, só aparece o lema (véxase *reducir, tabuco, administrar...*), mentres que se lema e hiperlema diverxen, recóllense ambos separados por unha barra oblicua (*dictar / ditar; xusticia / xustiza*), aparecendo en primeiro lugar o lema e a continuación o hiperlema:



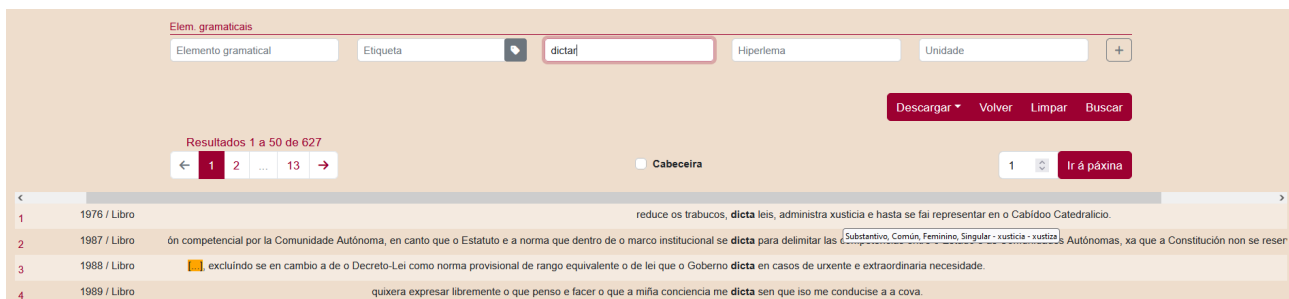
Fig. 30. Exemplo de lema e hiperlema diverxentes na análise en contexto.

Dende a versión 4.0, é posible visualizar tamén xa nas concordancias a análise de calquera elemento gramatical, non só o obxecto da procura, sen ter que acudir ao contexto: chega con colocar o rato enriba dun dos elementos, figura 31, para que emerxa unha caixa de texto coa análise:



The screenshot shows the search interface with 'dictar' entered in the search box. The results list shows several entries with the word 'dictar' highlighted in the text. A tooltip is visible over the word 'dictar' in the third result, showing its grammatical analysis: (Verbo, Presente, Indicativo, Terceira, Non aplica, Singular - dictar - ditar).

Fig. 31. Exemplo da caracterización dun elemento obxecto de busca.



This screenshot is similar to the previous one, but the tooltip is over the word 'xusticia' in the second result. The tooltip shows its grammatical analysis: (Substantivo, Común, Feminino, Singular - xusticia - xustizas).

Fig. 32. Exemplo da caracterización do elemento *xusticia* da concordancia 1 na xanela emerxente.

Como sinalamos máis arriba, o hiperlema asígnase só a partir da información que figura no lexicón e este é un recurso dinámico, non pechado. Como pode saber, polo tanto, quen acceda ao CORGA que hiperlemas se recoñecen e que lemas acubillan? Con esta finalidade, baixo a pestana **Guía** inclúese o ítem *Relación de hiperlemas*, que conduce á descarga dun arquivo en formato de só texto no que se recollen os hiperlemas que contén o lexicón. A listaxe ofrécese ordenada alfabeticamente polo hiperlema e cada liña contén un hiperlema, o lema ou un dos lemas que se agrupa baixo ese hiperlema e mais a clase de palabra á que se remite o lema. Os campos sepáranse con tabuladores e o hiperlema repítese tantas veces como pares “lema-clase de palabra” diferentes acolla. A figura seguinte exemplifica este formato:

```

veciño› veciño› Adxectivo↓
veciño› veciño› Substantivo↓
veciño› viciño› Adxectivo↓
veciño› viciño› Substantivo↓
vector› vector› Substantivo↓
vectorial› vectorial› Adxectivo↓
vectorización› vectorización› Substantivo↓
veda› veda› Substantivo↓
vedar› vedar› Verbo↓
vedete› vedete› Substantivo↓

```

Fig. 33. Mostra da visualización do arquivo de hiperlemas que contén o lexicón.

Cómpre advertir así mesmo que cando a identificación dunha forma dada se debe ao corpus de adestramento, a análise que para ela ofrezca o corpus etiquetado automaticamente achegará lema, mais non hiperlema, nin recoñecerá todos os elementos do paradigma, senón só aqueles que se

documenten no subcorpus de adestramento. É o que ocorre, por exemplo, naqueles casos en que identifica un lema porque aparece no corpus de adestramento, como sucede coa variante do indefinido *calquer*, non introducida ata o de agora no lexicón e polo tanto non acubillada aínda baixo o hiperlema *calquera*. Así pois, á hora de realizar as buscas ou analizar os resultados é relevante telo presente.

Por último, na pestana **Frecuencias**, a semellanza dos datos que ofrece o CORGA para os lemas, elementos gramaticais ou palabras ortográficas, inclúense tamén as frecuencias para os hiperlemas. Neste punto é precisa outra puntualización: non debe confundirse o arquivo de hiperlemas que se proporciona na **Guía** (información extraída da base de datos léxica coa que traballa o etiquetador automático) coa información de frecuencias que se pode obter na pestana **Frecuencias** (de entre os hiperlemas incorporados ao lexicón, nas frecuencias só aparecen os que se documentan nos textos do corpus).

5.6.1 O hiperlema nos casos de lematización automática

Co fin de mellorar as posibilidades de recuperación de información, probamos unha nova estratexia para propiciar a anotación morfosintáctica das formas descoñecidas sen ter que proceder á súa introdución na base de datos léxica: a lematización mediante regras lingüísticas. Comezamos na versión 4.0 cos adverbios en *-mente*, os apreciativos en *-iño*, os elativos, as formas co prefixo *auto-* e unha parte das formas que presentan gheada:



The screenshot shows the search interface of the CORGA corpus. At the top, there is a search bar with the following fields: 'Elem. gramaticais', 'Elemento gramatical', 'Etiqueta', 'autoabastecer', 'Hiperlema', and 'Unidade'. Below the search bar are buttons for 'Descargar', 'Volver', 'Limpar', and 'Buscar'. The search results are displayed in a table with 10 rows, showing the year, source, and a snippet of text containing the word 'autoabastecer'. The results are: 1. 1984 / Revista: pero temos outro patrimonio escrito e documental que se autoabastece a diario, que medra e ten demandas. 2. 2013 / Revista: España autoabastecer se que o seu consumo en un ano, 7%, que o seu produto en un 0,2% e de... 3. 2011 / Libro: con tres faquires a o ano autoabastecemos nos. 4. 1989 / Libro: favoreceu a confección de roupas para os poboadores de as zonas costeiras, lonxe de centros urbanos, que precisaban autoabastecer se. 5. 1996 / Revista: a arte en o traballo teatral serve exclusivamente para autoabastecer se artisticamente o que fai ese traballo, pero para nada máis. 6. 2000 / Revista: ande parte de pequenos propietarios agrarios, propensos a capitalizar se para manter a súa independencia económica e a autoabastecer se de capital antes de acudir a o crédito bancario (un autoconsumo máis, c... 7. 2005 / Xornal: sorte de colonia agrícola en a badía Bodega, a o norte de San Francisco, con o cal contaban poder autoabastecer se de alimentos en o futuro. 8. 2007 / Xornal: distribuíron de momento ningún tipo de alimentos entre os afectados por las inundacións xa que até agora son capaces de autoabastecer se, segundo informaba onte o diario boliviano La Razón. 9. 2008 / Revista: as iniciativas que gañaron o certame, dicir que o primeiro proxecto consiste en un bioedificio que é quen de autoabastecer se a través de as enerxías renovables. 10. 2011 / Revista: os consumidores poderemos nos autoabastecer con a enerxía renovable que xeremos en a casa?

Fig. 34. Exemplo de recuperación favorecida pola lematización automática.

Na versión 4.1 establecemos as regras para a lematización automática das formas descoñecidas que presentan os prefixos *ex-*, *etno-*, *etno*, *macro*, *macro-*, *meta-*, *meta*, *micro-*, *micro*, *multi-*, *multi*, *tele-*, *tele*, *xeo-* e *xeo*, e mais implementamos as regras que facilitan o recoñecemento automático dos elementos que mostran na súa representación gráfica dous fenómenos fonéticos moi estendidos en galego: as formas con gheada (fóra os nomes propios) e mais as formas con seseo (fóra os nomes propios e aquelas unidades que coinciden con formas presentes no lexicón, en cuxo caso non dispoñemos aínda de recursos para poder formalizar a distinción).

Sen dúbida, o recoñecemento automático das formas con gheada e/ou seseo constitúe un fito no desenvolvemento do CORGA e abre as portas ó seu tratamento noutras ferramentas de procesamento da linguaxe natural. Vexamos unha mostra da súa anotación no corpus:

<p>* Documento: Culpable de asesinato Editorial: Edicións Xerais de Galicia Orixe: Escrita Subtipo: Novela</p>	<p>Autor: Heinze, Úrsula Data: 1993 Bloque: Ficción Áreas temáticas: Sen clasificar</p>	<p>Sexo-autor: Muller Medio: Libro Xénero: Narrativo Sección no texto: Corpo</p>
---	--	---

¡Como non!, respondeu o fillo no meu lugar.
 As mulleres aman moito cando aman.

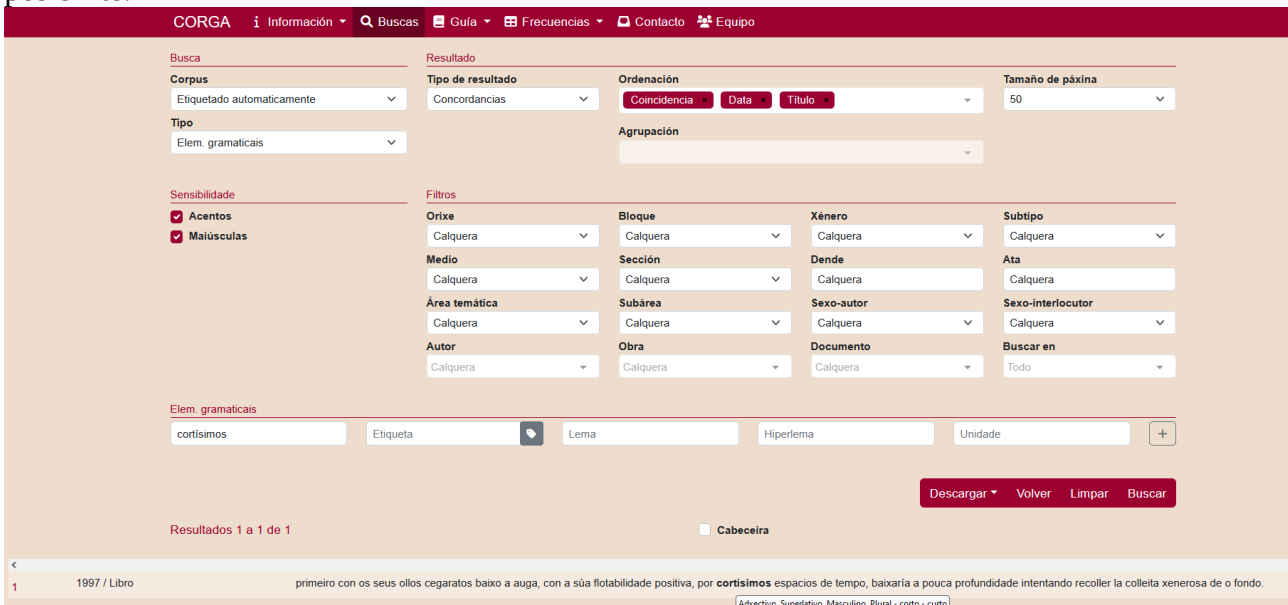
Parese que os homes casados exersen unha especial atracción sobre elas.

parese	que	os	homes	casados	exersen	unha	especial	atracción	sobre	elas	.
Vpi30s	Cs	Ddmp	Scmp	A0mp	Vpi30p	Difs	A0fs	Scfs	P	Rtp3fp	Q.
parecer	que	o	home	casado	exercer	un	especial	atracción	sobre	el	.

Fig. 35. Exemplo de anotación favorecida pola lematización automática de formas con gheada e seseo.

A delimitación do hiperlema nos casos de lematización automática require unha pequena explicación, pois a fiabilidade dos resultados empregando nas procuras o hiperlema vai diferir aquí. Así, asígnaselle lema dun xeito automático a calquera adverbio en *-mente* que apareza no corpus que non se rexistre no lexicón ou se documente no corpus de adestramento. O recoñecemento prodúcese a partir dunha regra que se basea na existencia da terminación en *-mente* e na exclusión dunha serie de formas que non son adverbios. Porén, neste caso o emprego de hiperlema nas procuras non é fiable, xa que se reproduce para el o lema. Isto explica que, por exemplo, non se recuperen mediante o hiperlema *claramente* as concordancias de *craramente*, inda que si se lematicen.

Para os apreciativos en *-iño*, os elativos, as formas con prefixos e as formas con gheada ou seseo, as regras de lematización automática fundaméntanse na presenza da forma base no lexicón; é dicir, esta ten que estar na base de datos léxica que manexa o etiquetador para poder ser recoñecida a forma derivada ou a variante con gheada e/ou seseo. Isto explica que as procuras por hiperlema sexan máis fiables. Así, *cortísimos* analízase como adxectivo en grao superlativo masculino plural (Amp) e lematízase automaticamente para *corto* grazas a unha regra lingüística. Logo, remítese ao hiperlema *curto* porque así consta na base de datos léxica, non porque se crease unha regra que o posibilite.



The screenshot shows the CORGA search interface. The search term is 'cortísimos'. The interface includes various filters such as 'Corpus', 'Tipo', 'Sensibilidade', 'Filtros', 'Ordenación', and 'Tamaño de páxina'. The search results show 'cortísimos' with a tag 'Lema' and a hyperlemma 'curto'. The interface also includes a search bar, a 'Cabeceira' checkbox, and a 'Descargar' button.

Fig. 36. Exemplo de lema e hiperlema favorecido pola lematización automática.

5.7. Inventario

As procuras na modalidade de **Elementos gramaticais** ofrecían na versión 3.2 a posibilidade de recuperar os datos das expresións coincidentes únicas mediante a selección do ítem respectivo no bloque **Resultado**, ben organizadas por elementos gramaticais, ben por lemas ou mesmo por hiperlemas. Esta opción transfórmase dende a versión 4.0 na modalidade **Inventario**, que se estende tamén ás buscas por palabra ortográfica e que vai integrar o ítem anteriormente denominado **Listaxes personalizadas**, o cal acollía un dicionario de frecuencias. Deste xeito, **Inventario** non só nos vai proporcionar, precisamente, o inventario de formas ortográficas, elementos gramaticais, lemas e hiperlemas que responden a un determinado criterio de procura, senón que engade a súa frecuencia e, se se precisa, tamén a súa distribución polos diferentes parámetros utilizados. Vemos o seu funcionamento cun exemplo.

Supoñamos que desexamos extraer as formas rematadas en *-metro* para comprobar os diversos aparellos de medida que se rexistran no corpus. Dado que os substantivos que esperamos recuperar presentan flexión de número, recorreremos ao lema para formular a consulta e cubrimos nese campo con **metro*. Deste xeito tiramos todas as palabras que figuran no corpus para as que consta un lema rematado en *-metro*, o que deixa fóra, por exemplo, os posibles casos de *cronometro*, primeira persoa do presente de indicativo de *cronometrar*. En total, o sistema devólvenos 14.956 casos, ordenados alfabeticamente por defecto polo termo obxecto da busca.

En versións anteriores á 3.2 do CORGA para obter unha lista coas formas únicas teríamos que descargar os resultados e logo traballalos nunha folla de cálculo ata reducilos ás formas distintas. Porén, dende a 3.2, seleccionando no bloque **Resultado** a opción de **Expresións coincidentes (lema)**, e agora **Inventario**, como amosan as figuras 37 e 38, esta tarefa xa a realiza a aplicación, que non só nos ofrece un caso por cada expresión coincidente, senón que mostra os datos relativos á súa frecuencia: número de casos que coinciden co recollido e número de documentos no que se rexistra.



Corpus de Referencia do Galego Actual

CORGA i Información B Buscas Guía Frecuencias Contacto Equipo

Busca

Corpus: Etiquetado automaticamente

Tipo: Elem. gramaticais

Sensibilidade: Acentos Maiúsculas

Resultado

Tipo de resultado: Inventario

Ordenación: Total Coincidencia

Tamaño de páxina: 50

Agrupación: Elemento gramatical

Filtros

Orixe: Calquera

Bloque: Calquera

Xénero: Calquera

Subtipo: Calquera

Medio: Calquera

Sección: Calquera

Dende: Calquera

Ata: Calquera

Área temática: Calquera

Subárea: Calquera

Sexo-autor: Calquera

Sexo-interlocutor: Calquera

Autor: Calquera

Obra: Calquera

Documento: Calquera

Buscar en: Todo

Elem. gramaticais

Elemento gramatical: Etiqueta

*metro

Hiperlema

Unidade

Descargar Volver Limpar Buscar

Fig. 37. Captación de datos para a visualización por *Inventario*.

Elem. gramaticais

Elemento gramatical: Etiqueta: *metro Hiperlema: Unidade:

Descargar Volver Limpar Buscar

Resultados 1 a 50 de 190(14.956)

← 1 2 ... 4 →

1 Ir á páxina

Selección de columnas

Valor absoluto	Lustro	Área temática
Medio	Orixe	Bloque
Subtipo	Xénero	Sexo-autor
Sexo-interlocutor		

		Total (54.737.447 / 57.693)
1	metros	5360 / 2361
2	quilómetros	3693 / 2032
3	metro	1138 / 618
4	parámetros	799 / 460
5	centímetros	793 / 430
6	diámetro	507 / 254
7	quilómetro	439 / 320
8	perímetro	300 / 158
9	milímetros	273 / 170
10	termómetro	211 / 106
11	centímetro	151 / 104
12	barómetro	116 / 63
13	parámetro	115 / 80

Fig. 38. Mostra da visualización de elementos gramaticais por *Inventario*.

Por defecto, os resultados só ofrecen o inventario de elementos gramaticais ou de formas ortográficas, en función da modalidade escollida, mais esta utilidade é sumamente flexiva como imos ver deseguido. No ítem **Agrupación**, que se activa tras seleccionar ver os resultados mediante a opción **Inventario**, o usuario pode seleccionar ademais a clase de palabra, a etiqueta completa, o lema, o hiperlema e/ou a unidade ortográfica, de xeito que os resultados se agruparán atendendo aos parámetros seleccionados para cada procura (na figura 39, lema, o que nos permite comprobar que os casos antes desagregados en *quilómetros* e *quilómetro* con número de orde 2 e 7 da figura 38, se agrupan agora no resultado 2, *quilómetro*):

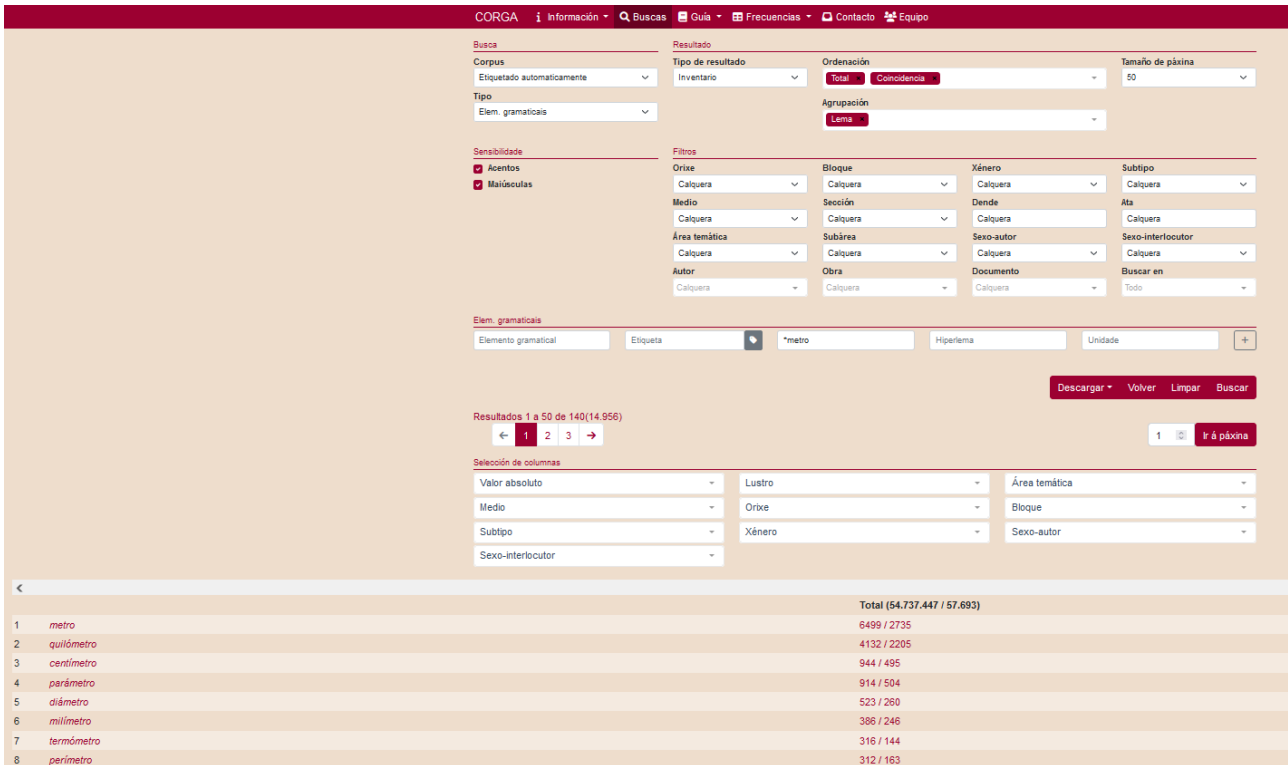


Fig. 39. Mostra da visualización de lemas por *Inventario*.

Ao comezo dos resultados, xusto antes de chegar aos resultados concretos, a aplicación de consulta permite ver ao momento, mediante a selección dos valores específicos nas columnas correspondentes e sen ter que premer de novo en **Buscar**, a distribución dos resultados obtidos polos parámetros clasificatorios que se precisen ou só por algún dos valores destes. É aí onde seleccionamos a frecuencia relativa e mais a distribución por xénero dos elementos gramaticais cuxo lema remata en **metro* e que pedimos agrupar tendo en conta a súa etiqueta (o que nos proporciona casos desagregados para as formas singulares e plurais), e mais o seu lema (obtemos un resultado específico para *quilómetro* e outro para *kilómetro*, por exemplo, que se agruparían en caso de organizalos por *hiperlema*). Vémolo na figura 40:

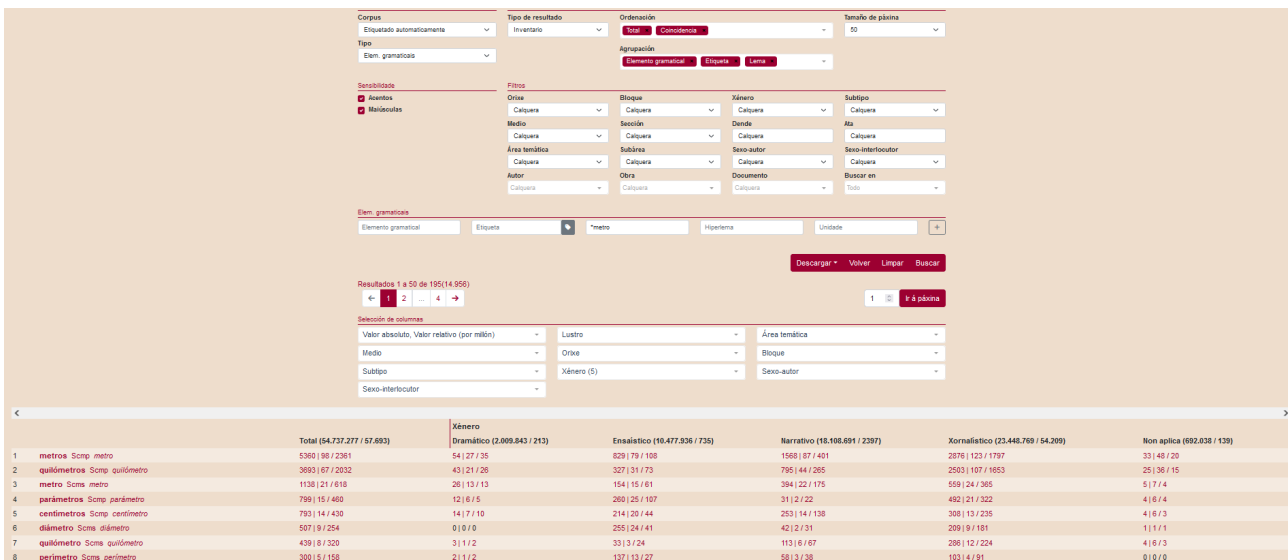


Fig. 40. Mostra da visualización por *Inventario* e distribución dos elementos por xénero.

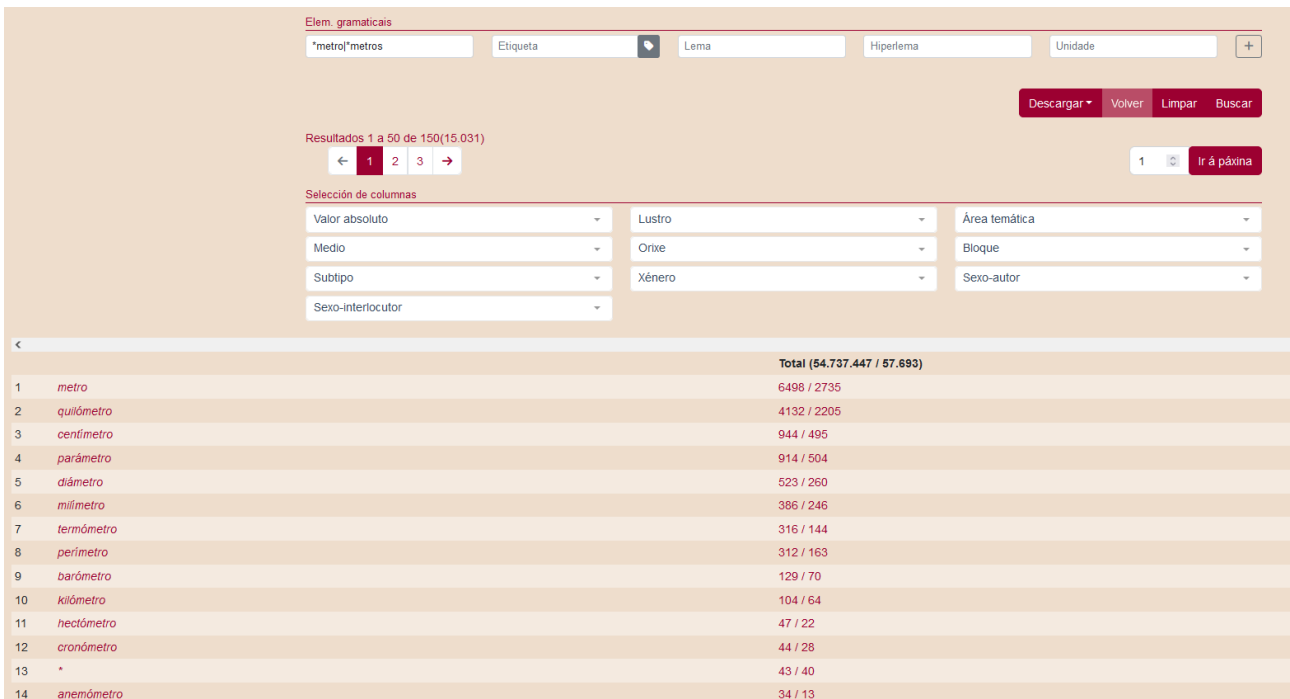
Premendo nalgún dos números destacados en vermello accédese ás concordancias do resultado seleccionado, por exemplo, para *parámetros*, as catro concordancias para o número 4 da columna *Non aplica*, correspondente a textos de procedencia oral, os cales non se clasifican por xénero:



The screenshot shows the search interface with 'parámetros' in the search box. The results list shows four items, with the 4th item selected. A dropdown menu is open for this item, showing details like 'Medio: Audiovisual', 'Orixe: Oral', and 'Xénero: Non aplica'.

Fig. 41. Concordancias do resultado 5 para a columna *Non aplica* da consulta anterior.

Agora ben, as buscas por lema só devolven resultados se no sistema consta dito lema⁷, polo que para cubrir posibles carencias nun caso coma o que nos ocupa, antes de analizar os datos, é aconsellable completar as procuras realizándoas tamén mediante o campo *Elemento gramatical*. Para iso cubrimos con **metro|*metros* na caixa do elemento, de maneira que imos obter todas as unidades que aparecen no corpus rematadas en *-metro* ou *-metros*, e mantemos no tipo de resultado **Inventario**. A figura 42 permite que contrastemos esta saída coa amosada na figura 39, en cuxa confrontación se percibe, agora si, un resultado para todas as ocorrencias que no corpus rematan en *-metro* ou *-metros*, incluídas as que non teñen lema asignado, circunstancia esta última representada no posto 13 por un asterisco (con premer no asterisco vemos que formas concretas se acubillan aí):



The screenshot shows the search interface with '*metro|*metros' in the search box. The results are displayed as a table with columns for the lemma and its frequency. The table is sorted by total frequency, with 'metro' having the highest count (6498 / 2735).

	Total (54.737.447 / 57.693)
1 metro	6498 / 2735
2 quilómetro	4132 / 2205
3 centímetro	944 / 495
4 parámetro	914 / 504
5 diámetro	523 / 260
6 milímetro	386 / 246
7 termómetro	316 / 144
8 perímetro	312 / 163
9 barómetro	129 / 70
10 kilómetro	104 / 64
11 hectómetro	47 / 22
12 cronómetro	44 / 28
13 *	43 / 40
14 anemómetro	34 / 13

Fig. 42. Mostra da visualización dos elementos rematados en *-metro* ou *-metros* agrupados polo lema.

Se acudimos agora ás concordancias do resultado número 13, premendo en **43 / 40**, ou sexa, clicando nos datos sobre o número de ocorrencias e documentos onde se localizan formas rematadas

7 Fóra os casos de lematización automática que vimos na epígrafe anterior.

en *-metro* ou *-metros* que non teñen lema, constatamos a presenza de unidades de medida (*newton/metro*, *céntimos/quilómetro*, *euro-viaxeiro-quilómetro*), compostos (*baremo-termómetro*, *limón-cronómetro*...) ou erros ortográficos (*quilometros*, *centrimetros*...), ademais das variantes, *animómetro* ou *cinemómetro*, formas presentes no CORGA mais non no lexicón.

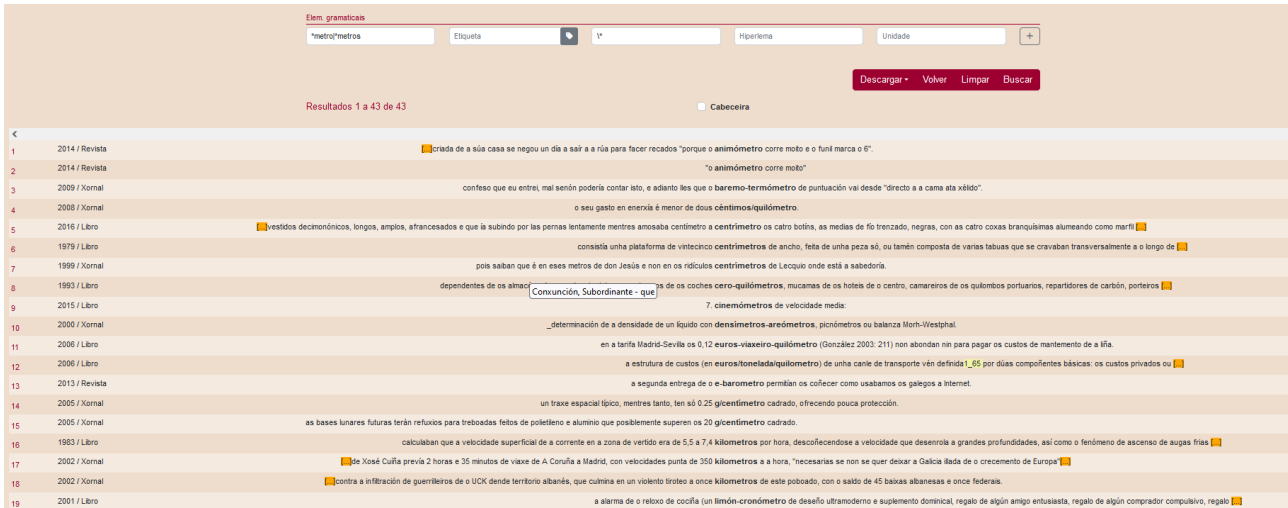
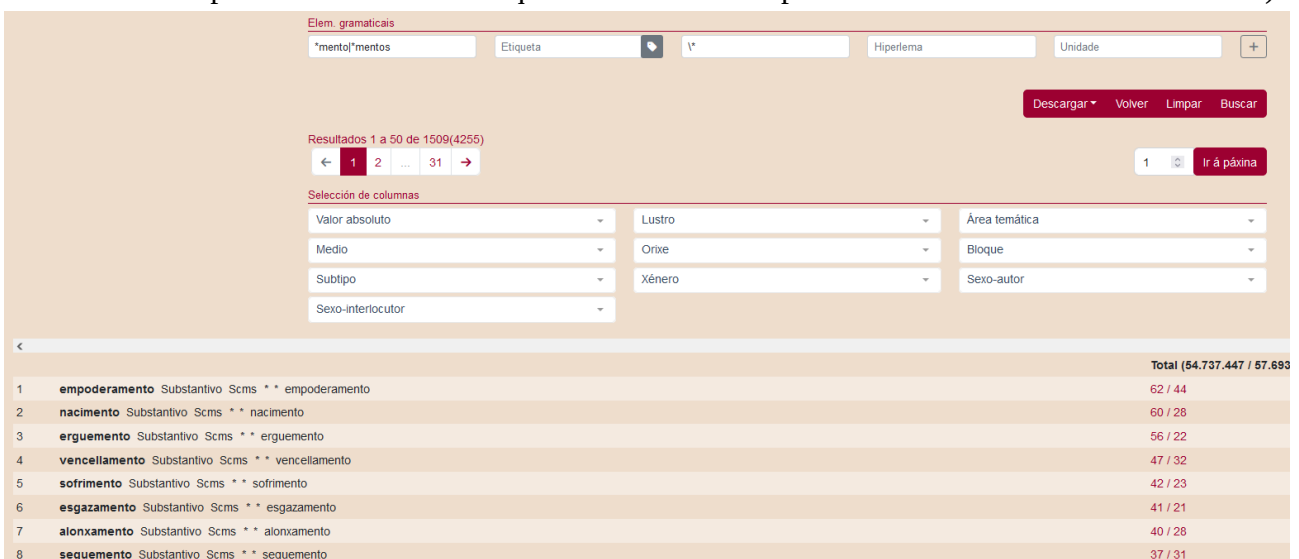


Fig. 43. Concordancias do resultado 13 do *Inventario* da consulta anterior.

Porén, posto que na consulta agrupamos os elementos gramaticais, premendo directamente no asterisco ‘ * ’, podemos mesmo visualizar o inventario dos elementos gramaticais, agrupados novamente en formas únicas, coas frecuencias que corresponden por elemento distinto, e podendo pedir ademais a súa distribución por calquera dos parámetros clasificatorios sen ter que realizar novas buscas. A información que se proporciona por defecto e nesta orde, exemplificada agora na Figura 44 para as formas rematadas en *-mento* ou *-mentos*, é a de elemento gramatical, categoría, etiqueta (emerxe ventá aclaratoria se se sitúa o rato enriba dela), lema (dado que non consta no lexicón nin é un dos sufixos lematizados automaticamente, no sitio do lema aparece un asterisco), hiperlema (mesma explicación que para o lema) e, por último, a forma ortográfica tal cal aparece no documento orixinal (péñese na súa importancia no caso de que os resultados correspondesen a formas verbais: poderíase discriminar aquí entre formas con pronomes enclíticos e formas sen eles).



Elemento gramatical	Categoría	Etiqueta	Lema	Hiperlema	Unicidade
empoderamento	Substantivo	Scms	**	empoderamento	62 / 44
nacimiento	Substantivo	Scms	**	nacimiento	60 / 28
erguemento	Substantivo	Scms	**	erguemento	56 / 22
vencellamento	Substantivo	Scms	**	vencellamento	47 / 32
sofrimento	Substantivo	Scms	**	sofrimento	42 / 23
esgazamento	Substantivo	Scms	**	esgazamento	41 / 21
aloxamento	Substantivo	Scms	**	aloxamento	40 / 28
seguemento	Substantivo	Scms	**	seguemento	37 / 31

Fig. 44. Inventario dos elementos contidos baixo o resultado con * da consulta descrita *supra*.

5.8. Coaparicións

Dende a versión 4.0 proporciónase para a modalidade de consulta por elementos gramaticais en proximidade⁸ un primeiro achegamento ao que son as coaparicións, tamén denominadas colocacións, presentes no CORGA. Trátase de formas que se combinan co elemento buscado cunha frecuencia superior á que sería esperable. Non obstante, polo de agora os resultados non teñen en conta a frecuencia da coaparición en si, senón que achegan só un índice da combinación da base escolleita con calquera colocativo⁹ na posición e distancias que o usuario determine:



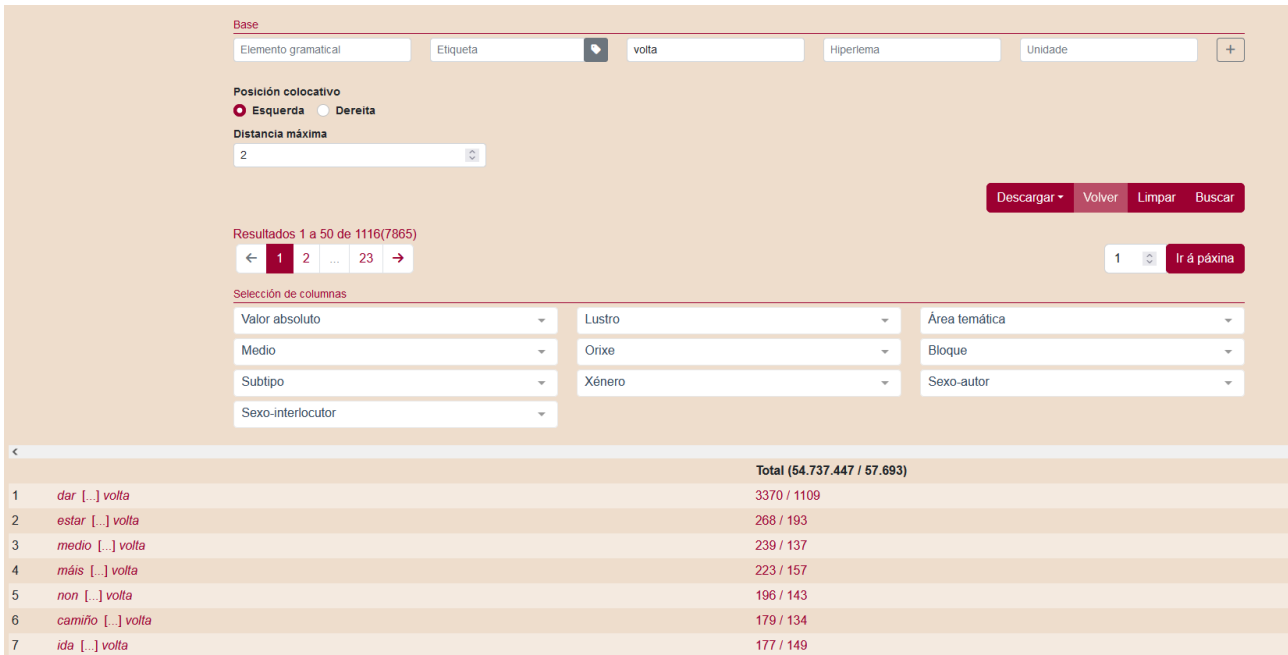
The screenshot shows the 'Corpus de Referencia do Galego Actual' search interface. The main navigation bar includes 'CORGA', 'Información', 'Buscas', 'Guía', 'Frecuencias', 'Contacto', and 'Equipo'. The search area is divided into several sections:

- Busca:** 'Corpus' (Etiquetado automaticamente), 'Tipo' (Elem. gramaticais próx.), 'Sensibilidade' (Acentos, Maiúsculas).
- Resultado:** 'Tipo de resultado' (Coaparicións), 'Ordenación' (Total, Coincidencia), 'Agrupación' (Lema), 'Tamaño de páxina' (50).
- Filtros:** A grid of filters including Orixe, Bloque, Xénero, Subtipo, Medio, Sección, Dende, Ata, Área temática, Subárea, Sexo-autor, Sexo-interlocutor, Autor, Obra, Documento, and Buscar en.
- Base:** 'Elemento gramatical' (Etiqueta), search term 'volta', 'Hiperlema', 'Unidade'.
- Posición colocativo:** Radio buttons for 'Esquerda' (selected) and 'Dereita'.
- Distancia máxima:** Input field with value '2'.

Fig. 45. Pantalla de captación de datos para *Coaparicións*.

Os filtros que se poden empregar para restrinxir as buscas son exactamente os mesmos dos que se dispón para obter as **Frecuencias**, as **Concordancias** ou o **Inventario**: parámetros de clasificación dos documentos, áreas temáticas, período cronolóxico, sección concreta do texto etc. Así, na figura anterior solicitamos as coaparicións para o lema *volta* de colocativos que se sitúen á esquerda da base e que estean a unha distancia de ata 2 posicións. Os resultados, organizados/agrupados por defecto polo lema, amósannos as combinacións obtidas xunto coas súas frecuencias. Comprobamos así que a base *volta* combínase con maior asiduidade, figura 46, con *dar*, *estar* e *medio*:

- 8 Os comodíns e a posibilidade de combinar calquera elemento, lema e unidade cobren o que sería a procura mediante a modalidade de palabra ortográfica próxima, polo que é innecesaria esa alternativa para as *Coaparicións*.
- 9 Só se teñen en conta como *colocativo* os elementos das clases de palabra abertas, ou sexa, substantivos, adxectivos, verbos e adverbios, excluindo destes últimos ademais os exclamativos-interrogativos e os relativos.



Base

Elemento gramatical: Etiqueta: volta Hiperlema: Unidade: +

Posición colocativo: Esquerda Dereita

Distancia máxima: 2

Resultados 1 a 50 de 1116(7865)

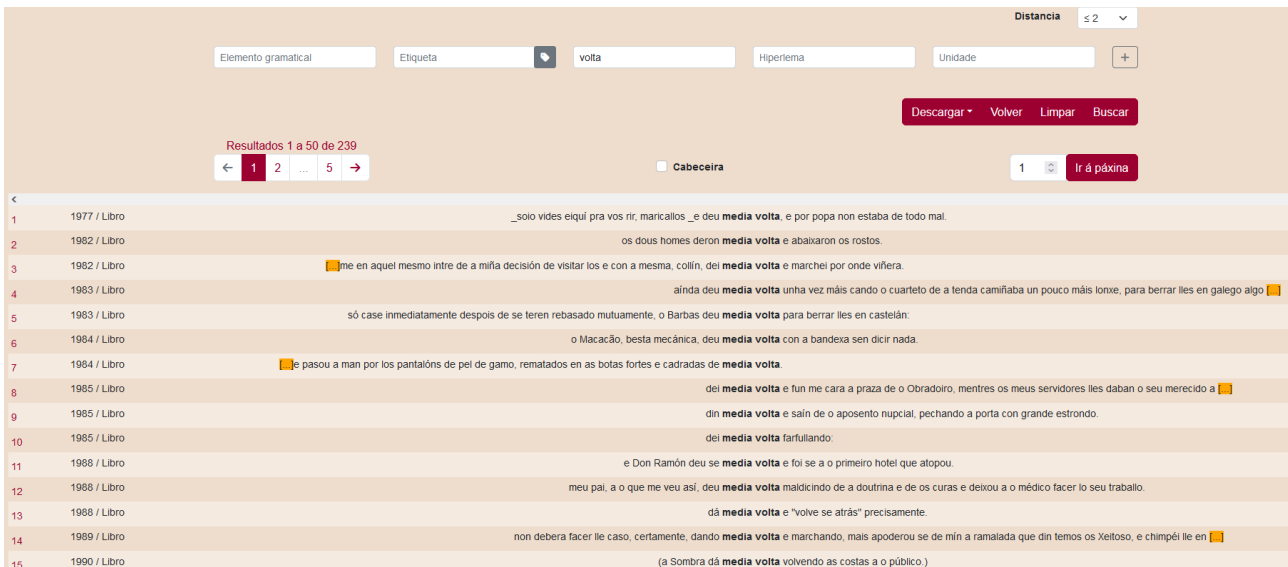
Selección de columnas:

Valor absoluto	Lustro	Área temática
Medio	Orixe	Bloque
Subtipo	Xénero	Sexo-autor
Sexo-interlocutor		

	Total (54.737.447 / 57.693)
1 dar [...] volta	3370 / 1109
2 estar [...] volta	268 / 193
3 medio [...] volta	239 / 137
4 máis [...] volta	223 / 157
5 non [...] volta	196 / 143
6 camiño [...] volta	179 / 134
7 ida [...] volta	177 / 149

Fig. 46. Pantalla de resultados de *Coaparicións*.

Dende esta pantalla podemos acceder directamente ás concordancias, por exemplo da combinación con *medio*, onde se constata que case na súa totalidade os casos corresponden a *media volta*, premendo na ligazón que subxace nas cifras situadas á dereita, figura 47:



Distancia: ≤ 2

Elemento gramatical: Etiqueta: volta Hiperlema: Unidade: +

Resultados 1 a 50 de 239

Cabeceira

1	1977 / Libro	_solo vides eiquí pra vos rir, maricallos_ e deu media volta , e por popa non estaba de todo mal.
2	1982 / Libro	os dous homes deron media volta e abaixaron os rostos.
3	1982 / Libro	me en aquel mesmo intre de a miña decisión de visitar los e con a mesma, collín, dei media volta e marchei por onde viñera.
4	1983 / Libro	aínda deu media volta unha vez máis cando o cuarteto de a tenda camiñaba un pouco máis lonxe, para berrar lles en galego algo
5	1983 / Libro	só case inmediatamente despois de se teren rebasado mutuamente, o Barbas deu media volta para berrar lles en castelán:
6	1984 / Libro	o Macacão, besta mecánica, deu media volta con a bandexa sen dicir nada.
7	1984 / Libro	pasou a man por los pantalóns de pel de gamo, rematados en as botas fortes e cadradas de media volta .
8	1985 / Libro	dei media volta e fun me cara a praza de o Obradoiro, mentres os meus servidores lles daban o seu merecido a
9	1985 / Libro	din media volta e saíu de o aposento nupcial, pechando a porta con grande estrondo.
10	1985 / Libro	dei media volta farfollando:
11	1988 / Libro	e Don Ramón deu se media volta e foi se a o primeiro hotel que atopou.
12	1988 / Libro	meu pai, a o que me veu así, deu media volta maldicindo de a doutrina e de os curas e deixou a o médico facer lo seu traballo.
13	1988 / Libro	dá media volta e "volve se atrás" precisamente.
14	1989 / Libro	non debera facer lle caso, certamente, dando media volta e marchando, mais apoderou se de mín a ramalada que din temos os Xeltoso, e chimpéi lle en
15	1990 / Libro	(a Sombra dá media volta volvendo as costas a o público.)

Fig. 47. Pantalla de concordancias á que se accede dende os resultados da busca por *Coaparicións*.

Así mesmo, podemos acceder ás combinacións de elementos gramaticais que se acubillan baixo o modelo representado por *dar [...] volta*, figura 48, premendo nesta ocasión sobre a propia combinación situada á esquerda da pantalla. Aquí lístanse, no modo **Inventario**, os 183 *types* que acollen as 3370 concordancias nas que aparece *dar [...] volta* no corpus:

Resultados 1 a 50 de 183(3370)

← 1 2 ... 4 → 1 Ir á páxina

Selección de columnas

Valor absoluto	Lustro	Área temática
Medio	Orixe	Bloque
Subtipo	Xénero	Sexo-autor
Sexo-interlocutor		

		Total (54.737.447 / 57.693)
1	dar Verbo V0f000 dar dar dar [...] volta Substantivo Scfs volta volta volta	589 / 353
2	dar Verbo V0f000 dar dar dar [...] voltas Substantivo Scfp volta volta voltas	349 / 234
3	deu Verbo Vei30s dar dar deu [...] volta Substantivo Scfs volta volta volta	312 / 181
4	dando Verbo V0x000 dar dar dando [...] voltas Substantivo Scfp volta volta voltas	269 / 191
5	dá Verbo Vpi30s dar dar dá [...] volta Substantivo Scfs volta volta volta	199 / 119
6	dar Verbo V0f000 dar dar darle [...] voltas Substantivo Scfp volta volta voltas	125 / 101
7	daba Verbo Vii30s dar dar daba [...] voltas Substantivo Scfp volta volta voltas	107 / 73
8	dando Verbo V0x000 dar dar dándolle [...] voltas Substantivo Scfp volta volta voltas	105 / 86
9	dá Verbo Vpi30s dar dar dá [...] voltas Substantivo Scfp volta volta voltas	103 / 82
10	deu Verbo Vei30s dar dar Deu [...] volta Substantivo Scfs volta volta volta	89 / 70
11	dando Verbo V0x000 dar dar dando [...] volta Substantivo Scfs volta volta volta	85 / 68
12	dá Verbo Vpi30s dar dar Dá [...] volta Substantivo Scfs volta volta volta	54 / 33

Fig. 48. Relación de elementos aos que se accede dende un resultado da busca por *Coaparicións*.

A información que aparece por defecto para cada unha destas combinacións é a seguinte: a base e o colocativo destacados en negra e separados por [...], e para cada un dos elementos da coaparición proporcióname clase de palabra, etiqueta, lema, hiperlema e unidade ortográfica. Basta colocar o rato enriba dun segmento para que emerxa unha caixiña de texto coa información explicativa pertinente, ben coa identificación de se é elemento gramatical, lema, hiperlema ou unidade ortográfica, ben para desenvolver a etiqueta de xeito que sexa transparente para calquera persoa. Por exemplo, para Vei30s que aparece na terceira posición:

Descargar Volver Limpar Buscar

Resultados 1 a 50 de 183(3370)

← 1 2 ... 4 → 1 Ir á páxina

Selección de columnas

Valor absoluto	Lustro	Área temática
Medio	Orixe	Bloque
Subtipo	Xénero	Sexo-autor
Sexo-interlocutor		

		Total (54.737.277 / 57.693)
1	dar Verbo V0f000 dar dar dar [...] volta Substantivo Scfs volta volta volta	589 / 353
2	dar Verbo V0f000 dar dar dar [...] voltas Substantivo Scfp volta volta voltas	349 / 234
3	deu Verbo Vei30s dar dar deu [...] volta Substantivo Scfs volta volta volta	312 / 181
4	dando Verbo <small>Etiqueta: Verbo, Pretérito, Indicativo, Terceira, Non aplica, Singular</small> Scfp volta volta voltas	269 / 191
5	dá Verbo Vpi30s dar dar dá [...] volta Substantivo Scfs volta volta volta	199 / 119

Fig. 49. Desenvolvemento dunha etiqueta no inventario dun resultado de busca por *Coaparicións*.

Naturalmente, tamén desde este punto se pode navegar polas concordancias de calquera das combinacións de *dar [...] volta* ou ver a súa distribución por calquera dos parámetros que figuran xusto antes dos resultados, sen necesidade de ter que realizar unha nova procura, só con seleccionar os valores que se desexen na columna que corresponda, por exemplo, para ver a distribución consonte a orixe do documento:

Descargar Volver Limpar Buscar

Resultados 1 a 50 de 183(3370)

1 2 4 ir á páxina

Selección de columnas

Valor absoluto Lusto Área temática
Medio Orixe (2) Bloque
Subtipo Xénero Sexo-autor
Sexo-interlocutor

		Total (54.737.277 / 57.693)	Orixe Escrita (54.045.239 / 57.554)	Oral (692.038 / 139)
1	dar Verbo VOI000 dar dar dar [...] volta Substantivo Scfs volta volta volta	589 / 353	578 / 345	11 / 8
2	dar Verbo VOI000 dar dar dar [...] voltas Substantivo Scfp volta volta voltas	349 / 234	348 / 233	1 / 1
3	deu Verbo Vei30s dar dar deu [...] volta Substantivo Scfs volta volta volta	312 / 181	311 / 180	1 / 1
4	dando Verbo VOx000 dar dar dando [...] voltas Substantivo Scfp volta volta voltas	269 / 191	266 / 188	3 / 3
5	dá Verbo Vpi30s dar dar dá [...] volta Substantivo Scfs volta volta volta	199 / 119	197 / 117	2 / 2
6	dar Verbo VOI000 dar dar darlle [...] voltas Substantivo Scfp volta volta voltas	125 / 101	122 / 98	3 / 3
7	daba Verbo Vii30s dar dar daba [...] voltas Substantivo Scfp volta volta voltas	107 / 73	107 / 73	0 / 0
8	dando Verbo VOI000 dar dar dándolle [...] voltas Substantivo Scfp volta volta voltas	105 / 86	102 / 84	3 / 2
9	dá Verbo Vpi30s dar dar dá [...] voltas Substantivo Scfp volta volta voltas	103 / 82	103 / 82	0 / 0

Fig. 50. Distribución da coaparición dar [...] volta segundo a orixe do documento.

5.9. Nómima

Dende a versión 4.0 volvemos recuperar a nómina de autores e obras que compoñen o corpus, de maneira que seleccionando no bloque **Resultado** o tipo **Nómima**, automaticamente desaparecen as modalidades de consulta por palabras ortográficas ou elementos gramaticais e no sitio aparece a opción **Documentos**. Pódese así pescudar que documentos constan no CORGA para un autor concreto, saber que documentos se introduciron para un subtipo determinado ou obter a listaxe dos documentos que son de autoría feminina, por exemplo. O resultado da procura ofrece os datos relativos á referencia da obra e mais os datos estatísticos relevantes:

Corpus de Referencia do Galego Actual

CORGA Información Buscas Guía Frecuencias Contacto Equipo

Baixa Resultado

Corpus Tipo de resultado Nómima Ordenación Tabo Tamaño de páxina 50

Tipo Documentos Agrupación

Sensibilidade

Alentidos
Máisculas

Filtros

Orixe Catquera Bloque Catquera Xénero Catquera Subtipo Catquera
Medio Sección Catquera Dende Alta
Área temática Subárea Catquera Sexo-autor Catquera Sexo-interlocutor Catquera
Autor Cabana, Darío Xohán Obra Catquera Documento Catquera Buscar en Todo

Volver Limpar Buscar

Resultados 1 a 9 de 9

Documento	Autor	Editorial	Data	Medio	Sección	Orixe	Bloque	Xénero	Subtipo	Área temática	Palab. ortográficas 44.323	Elem. gramaticais 54.026
A excursión	Cabana, Darío Xohán	Transportes Castromil	1992	Libro	Non aplica	Escrita	Ficción	Narrativo	Relato curto	Sen clasificar	3193	3850
A illa	Cabana, Darío Xohán	Nigra	1994	Libro	Non aplica	Escrita	Ficción	Narrativo	Relato curto	Sen clasificar	1401	1815
Cerco de ferro	Cabana, Darío Xohán	Nigra	1994	Libro	Non aplica	Escrita	Ficción	Narrativo	Relato curto	Sen clasificar	1501	1917
O caldeiro	Cabana, Darío Xohán	Nigra	1994	Libro	Non aplica	Escrita	Ficción	Narrativo	Relato curto	Sen clasificar	5812	7121
O foxidor	Cabana, Darío Xohán	Edicións Xerais de Galicia	1994	Libro	Non aplica	Escrita	Ficción	Narrativo	Relato curto	Sen clasificar	9178	11.156
A Maniñeira de Qulmas	Cabana, Darío Xohán	Edicións Xerais de Galicia	1994	Libro	Non aplica	Escrita	Ficción	Narrativo	Relato curto	Sen clasificar	2462	3011
Rubén Cruzoi	Cabana, Darío Xohán	Edicións Xerais de Galicia	1994	Libro	Non aplica	Escrita	Ficción	Narrativo	Relato curto	Sen clasificar	11.432	13.757
O Vello da Montaña	Cabana, Darío Xohán	Edicións Xerais de Galicia	1994	Libro	Non aplica	Escrita	Ficción	Narrativo	Relato curto	Sen clasificar	3148	3814
Severo da Melisenda	Cabana, Darío Xohán	Edicións Xerais de Galicia	1994	Libro	Non aplica	Escrita	Ficción	Narrativo	Relato curto	Sen clasificar	6196	7585

Fig. 51. Resultados da Nómima para o autor Darío Xohán Cabana.

6. Filtros

A maiores dos distintos tipos de consulta que vimos de describir, o sistema permite a recuperación de datos sobre a totalidade do corpus ou ben sobre o subcorpus virtual creado directamente por quen fai a consulta en función dos diferentes parámetros utilizados na codificación e estruturación dos textos. Isto é,

- Período cronolóxico

- Parte estrutural
- Medio
- Sección
- Área e subáreas temáticas
- Autor
- Sexo do autor
- Sexo do interlocutor
- Obra
- Documento
- Clasificación textual

Deseguido imos describir brevemente cada un deles, comezando polos filtros que se manteñen inalterados ou coa adición de novos valores con respecto a versións anteriores, para rematar coa descrición daqueles nos que se producen os maiores cambios.

6.1. Período cronolóxico

O usuario pode realizar a consulta sobre todo o corpus, e polo tanto en todo o período cronolóxico que este abrangue (dende 1975 ata a actualidade), ou pode restrinxir a consulta a un ano concreto ou ao rango temporal que lle interese determinar. Malia que o corpus se estende ata a actualidade, a data límite ata a que se pode buscar é nesta versión 2021, pois non se introduciron polo momento documentos datados con posterioridade a ese ano. Para seleccionar un período diferente do total, só hai que premer na pestana **Dende** e escoller entre todos os anos aquel dende o que se quere iniciar o período, para logo premer na pestana **Ata** e marcar o ano desexado de remate.

6.2. Parte estrutural do documento

Calquera dos tipos de busca que vimos de ver na descrición da aplicación pode efectuarse sobre os textos completos ou sobre unha parte estrutural concreta do documento (por exemplo, os titulares de novas xornalísticas). Por defecto realízanse as buscas en todo o documento. Para seleccionar unha parte concreta débese premer na pestana **Buscar en** e alí activar as que interesen. As posibilidades son as seguintes, todas elas transparentes: *acoutación, alongamento, apéndice, cita, corpo, dedicatoria, encabezamento, interlocutor, nota, pé de foto, prólogo, resumo, solapado, texto riscado, texto unido, titular e transcripción dubidosa*.

Interlocutor asigna o texto a un personaxe concreto dunha obra de teatro ou guión, así coma a un interveniente nunha entrevista, coloquio, mesa redonda etc., ou o locutor dun informativo, un participante nun faladoiro ou calquera tipo de falante nunha transcripción. É dicir, *interlocutor* localiza o fragmento dialogado emitido por un emisor do que se proporciona o nome explicitamente no documento e que é eliminado do texto e convertido en marca no documento que se incorpora ao corpus, co que se evita que os nomes dos interlocutores terxiversen os datos léxicos do galego interferindo nas frecuencias e, á vez, delimita unha parte grande da oralidade na escrita.

Dende a versión 4.1 incorpórase ademais a posibilidade de buscar en texto marcado con alongamento (tanto para o rexistro oral coma para o escrito), ou que aparece en solapamentos; tamén se dá a opción de recuperar aquelas formas que presentan unha transcripción dubidosa, así como aquelas formas que se presentan no texto riscadas, a modo de relación paradigmática co texto

que as substitúe, ou palabras que agora se visualizan nunha grafía convencional mais que aparecían no texto orixinal unidas con guión ou se soletreaban.

Debe terse en consideración que se existe contradición entre o **Tipo de texto** ou o **Medio** escolleito e mais a parte do texto na que se pretende buscar, por exemplo se se selecciona **Libro** en **Medio** e en **Buscar en** se opta por *titular*, o sistema non devolverá resultados.

6.3. Medio

A incorporación de transcricións de programas radiofónicos e de guións de series televisivas, así como de blogs, esixiu ampliar os medios anteriormente dispoñibles (**Xornal**, **Revista** e **Libro**) cos novos valores **Audiovisual** e **Internet**.

6.4. Sección

As noticias procedentes de xornais clasifícanse segundo a sección na que aparecían, o que favorece que na recuperación de información se poidan facer consultas tendo en conta este parámetro. As seccións existentes son: *Actualidade, Área de Compostela, Campus, Cultura e Sociedade, Deportes, Economía, España, Galicia, Internacional, Opinión, Suplemento, Tempo e TV*. Os valores que se seleccionen para os campos **Tipo de texto** e **Medio** condicionan a activación/desactivación da **Sección**, de xeito que esta se activa cando as consultas se lanzan sobre un conxunto de textos que inclúa os xornais, e se deshabilita en caso contrario.

Dado que a agrupación natural das diferentes noticias nos xornais é por seccións, no momento de incorporar os xornais ao corpus pareceunos atinado conservar a devandita clasificación, posto que esta pode repercutir en diferenzas léxicas.

6.5. Área temática

Todas as noticias de xornais e revistas, as entradas dos blogs, os diversos tipos de textos ensaísticos ou os prólogos introdutorios asinados por un autor diferente do da colección ou obra caracterízanse tematicamente con ata tres áreas temáticas, o que permite no sistema de recuperación de información empregar este parámetro para restrinxir as ocorrencias da forma ou estrutura que se procura á área e/ou subárea especificadas.

Deste xeito, por exemplo, se escribimos *tableta** no campo de texto **Palabras ortográficas** e restrinximos a **área temática** ao valor *Ciencias e tecnoloxía*, especificando que o valor para a **subárea temática** é *Tecnoloxía e industria*, obtemos a información de que as formas *tableta* e *tabletas*, que aparecen un total de 79 veces, concorren con *tablet* e *tablets*, mais de momento lonxe aínda da frecuencia de uso do anglicismo (130 ocorrencias).

Non obstante, cómpre ter presente que, malia que nos textos asignamos as áreas temáticas pola súa predominancia, esta non se traslada na indexación dos textos á versión en liña, ou sexa, as áreas temáticas non están xerarquizadas cando se recupera a información. Esta é a explicación, se se observan as frecuencias completas da procura que propoñiamos no parágrafo anterior, de que no cadro correspondente ás áreas temáticas se rexistren outras áreas á parte da que nos interesaba, como se mostra na figura seguinte, onde, non é que haxa máis de 79 casos, senón que para 42 deles, por exemplo, ademais da subárea esixida, o documento está clasificado tamén coa clave de economía e política.

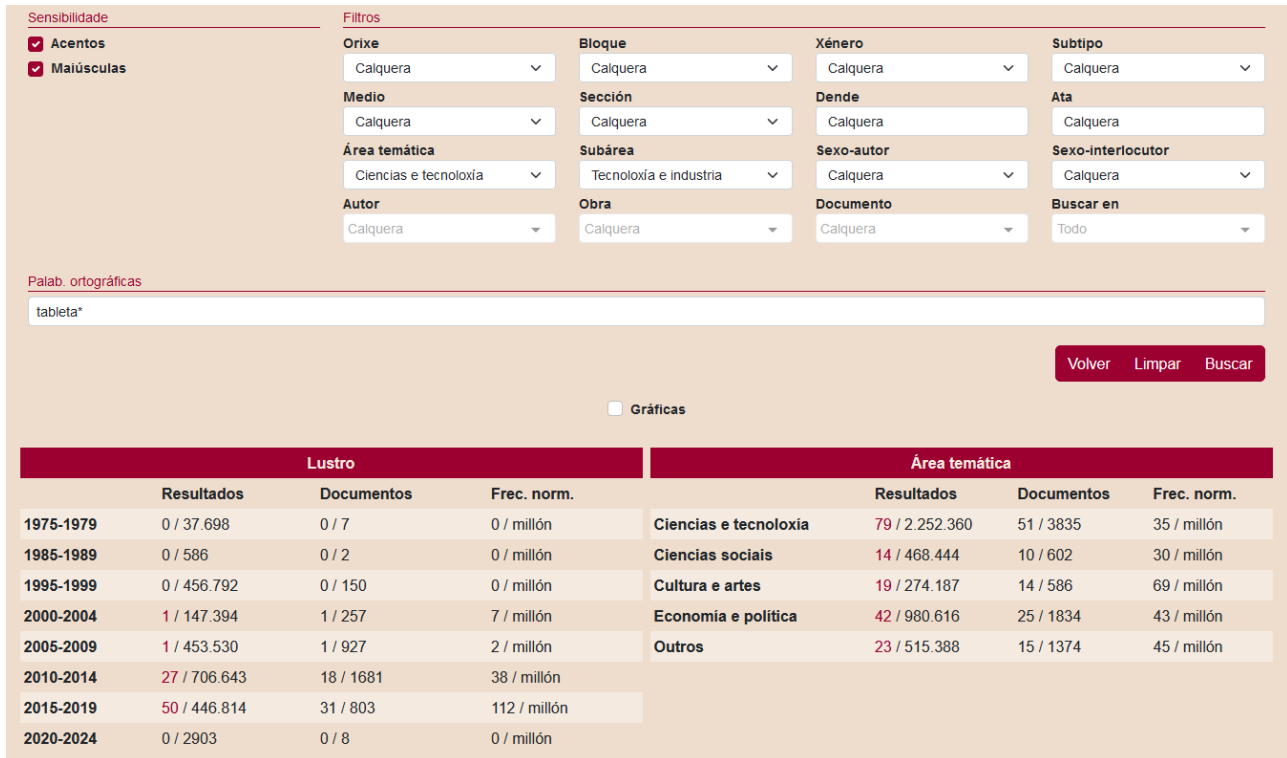


Fig. 52. Mostra da visualización das frecuencias por área temática.

As áreas e subáreas temáticas empregadas para a clasificación dos documentos comprendidos nos xéneros ensaístico e xornalístico son as que recolleemos na seguinte táboa:

ÁREAS TEMÁTICAS					
	Economía e política	Cultura e artes	Ciencias sociais	Ciencias e tecnoloxía	Outros
SUBÁREAS TEMÁTICAS	Política	Audiovisuais e espectáculo	Lingua	Sanidade	Deportes
	Desenvolvemento e infraestruturas	Medios de comunicación	Literatura	Bioloxía, botánica, ecoloxía, zooloxía e paleontoloxía	Turismo
	Emprego, traballo, industria	Artes gráficas e plásticas	Relixión	Tecnoloxía e industria	Afeccións e asuntos domésticos
	Sector servizos	Patrimonio, arquitectura, arquivos	Historia e xeografía	Medio, astronomía e xeoloxía	Actualidade, sucesos, homenaxes, inauguracións
	Explotación primaria		Civilización, etnoloxía, arqueoloxía e antropoloxía	Matemáticas e estatística	Biografía
	Economía, facenda, bolsa		Pensamento, ética e filosofía	Química, bioquímica e farmacia	Nota prologal
	Ordenación sanitaria		Socioloxía e psicoloxía		
	Xustiza, lexislación, dereito		Erotismo e sexoloxía		
Asuntos sociais		Astroloxía e ocultismo			

	Ordenación académica
--	----------------------

Táboa 1. Valores das áreas e subáreas temáticas nas que se clasifican os textos do CORGA.

6.6. Autor

Dende a versión 4.0, o usuario pode realizar a consulta sobre un ou varios dos autores cuxas obras se integran no corpus mediante a selección do nome ou nomes desexados. O formato escolleito é o mesmo que aparece para o campo autor cando se visualizan as cabeceiras: *Apellido(s)*, *Nome*. Por defecto amósanse só os primeiros 1000 elementos da listaxe, polo que, se a autoría buscada non se atopa entre eles, debe introducirse algún valor no campo do selector do filtro. Por exemplo, ao escribir *Cano*, como se aprecia na parte inferior da figura seguinte, emerxen na relación as dúas autoras cuxo apelido comeza con esas letras:



The screenshot shows the CORGA search interface. The search bar contains 'Cano'. The 'Autor' filter dropdown is open, showing a list of suggestions: 'Lizcano, R.', 'Canosa, María', 'Canosa, Tamara', 'Lizcano, Rocío', 'Lezcano, Arturo', and 'Lezcano González, Arturo'. The interface includes various filters for 'Orixe', 'Medio', 'Área temática', 'Subárea', 'Obra', 'Bloque', 'Sección', 'Xénero', 'Dende', 'Sexo-autor', 'Documento', 'Subtipo', 'Ata', and 'Sexo-interlocutor'. The search results are currently empty.

Fig. 53. Exemplo de selección dunha autoría concreta.

Así mesmo, cómpre ter en conta que o metadato relativo á autoría dun documento mantén a grafía que figura en cada obra, sen regularizacións nin normalizacións de ningún tipo, de xeito que pode existir máis dunha variante. Por exemplo:

González Reigosa, Carlos

Reigosa, Carlos G.

6.7. Sexo do autor

Dende a versión 4.0 a aplicación de consulta introduce a posibilidade de realizar procuras tendo en conta o sexo do autor dos documentos. Os valores posibles para ese novo filtro son: *Ambos*, *Descoñecido*, *Home*, *Muller* e *Non aplica*. No momento en que se incorpore algún

documento cuxo autor explicita esa alternativa, estes valores ampliaranse coa opción *Non binario*.

O criterio que seguimos para aplicar o valor *sexo* no campo *Autor* é o do nome de pía do autor, de xeito que

Canosa, María => clasifícase co valor *muller*.

Méndez Ferrín, Xosé Luís => clasifícase co valor *home*.

Álvarez, Dani => caracterízase como *descoñecido*, pois o diminutivo pode corresponder tanto a *Daniel* como a *Daniela*.

Abreviaturas, Redacción, Axencia, Colectivos etc. => caracterízanse co valor *descoñecido*.

Esta clasificación susténtase en mellorar as posibilidades de obtención e clasificación da información atendendo a unha perspectiva de xénero sen consumir esforzos indagando sobre a que corresponden as abreviacións das autorías, moi claras no caso de Méndez Ferrín, X. L., mais problemáticas na maior parte dos casos. No xénero ensaístico, narrativo e dramático, sempre que o sabemos, asignamos o sexo do autor independentemente de que haxa abreviacións ou non nos nomes. Deste xeito tanto Méndez Ferrín, X. L. coma Jaureguizar terán atribuído en sexo do autor o valor *home*. Pola súa banda, nos xornais, dado que estamos ante documentos moito máis breves e cunhas características moi diferentes, senón casos moi claros en que coaparezan nome completo e nome con abreviación, se non se pode determinar o sexo do autor, asígnaselle o valor *descoñecido*.

Ademais dos valores *Descoñecido*, *Home* e *Muller*, o filtro por sexo do autor inclúe as alternativas *Ambos* e *Non aplica*. A asignación a un ou outro realízase como segue:

Ambos. Clasifica documentos de autoría coral na que participan autores de ambos os sexos.

Non aplica. Clasifica as transcripcións de audio, que se caracterizan por carecer dun autor xeral do documento e no que a autoría debe realizarse en función de quen emite cada unha das alocucións.

6.8. Sexo do interlocutor

Co propósito de facilitar a recuperación dos datos do rexistro oral tendo en conta o criterio *sexo* do falante, incorporouse na versión 4.1 un filtro máis á oferta xa existente: *sexo do interlocutor*, en paralelo ao *sexo do autor*, para, alén de favorecer o estudo da fala feminina ou masculina, facilitar os cruzamentos entre autores dun sexo dado e personaxes do sexo contrario.

Os criterios para a súa asignación son os mesmos que para o sexo do autor, coa única excepción dos relativos ao valor *non pertinente*, non preciso para o campo do sexo do autor. Na ficción é habitual dotar de voz personaxes non humanos e mesmo inanimados. O valor *non pertinente* aplícase precisamente nestes casos, en seres asexuados.

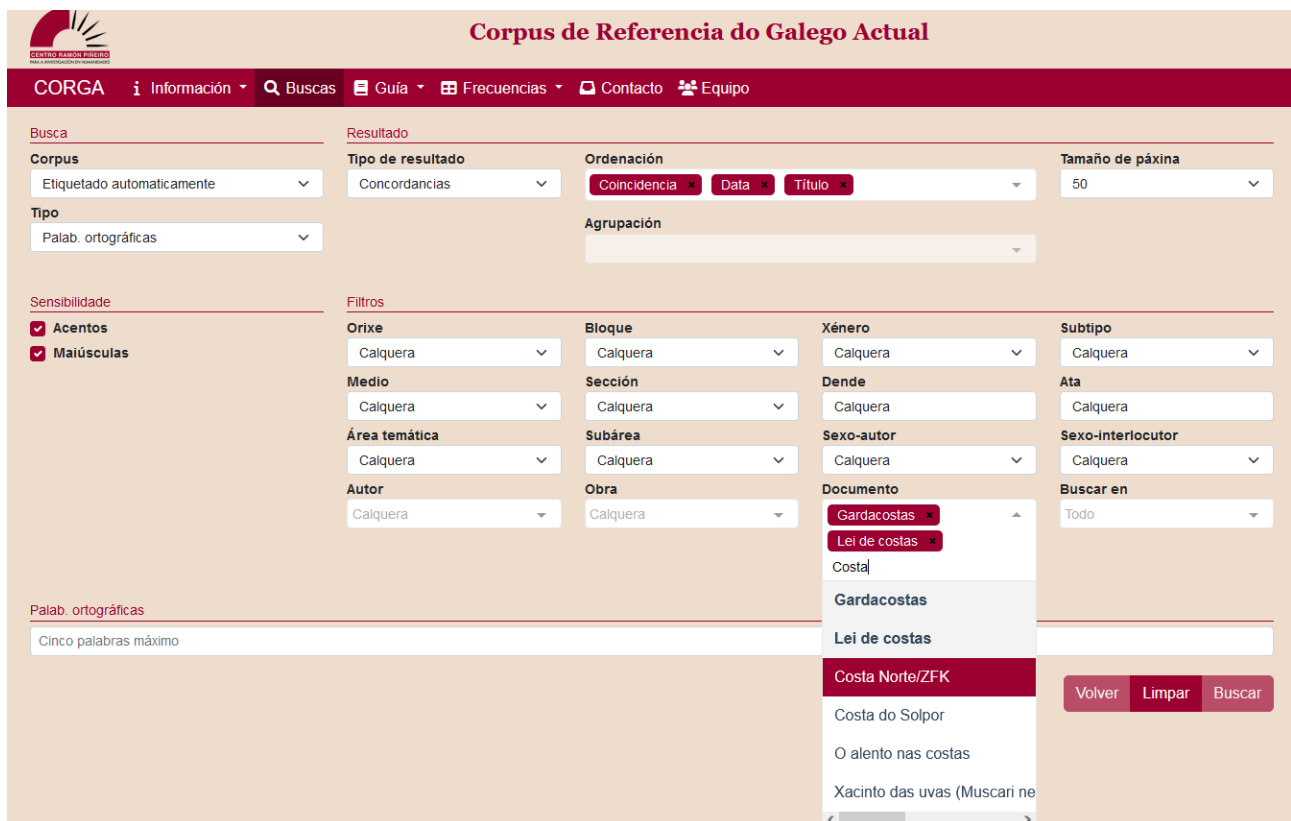
6.9. Documento

A identificación de documento susténtase no CORGA principalmente na súa autoría: a existencia de autores distintos –non confundir con autoría colectiva– dá lugar a documentos diferentes. Así, un xornal ou unha revista son textos que están organizados en múltiples noticias, cada unha delas responsabilidade dun autor concreto e clasificada tematicamente coas subáreas pertinentes. De xeito similar, unha colección de relatos ou unha obra de teatro adoita estar

prologada por un autor distinto do autor dos relatos ou peza teatral, inda que forme parte fisicamente do mesmo libro, o que os converte en documentos diferentes para o sistema, e en consecuencia leva asociada unha cabeceira na que se recollen os metadatos e se dá conta do aniñamento. O mesmo sucede con calquera obra que reúna textos de diversos autores, xa for obras de teatro, xa for relatos, xa for artigos científicos ou xornalísticos.

Así mesmo, nas coleccións de relatos, nas entradas de blogs e nas compilacións ensaísticas, sexan dun único autor ou pertenzan a varios, cada relato, cada entrada e cada artigo constitúe un documento diferente porque cada un deles presenta unicidade, representada esta nunha cabeceira cos metadatos específicos: título, autor, áreas temáticas etc., de aí o alto número de “documentos” que se observan nos datos das frecuencias do CORGA e na relación de documentos que integra o filtro **Documentos** na aplicación de consultas.

Ao igual que sucedía co filtro Autor, tamén aquí se amosan por defecto só os primeiros 1000 elementos da listaxe, polo que convén introducir algún valor no campo do selector do filtro para facilitar a selección. Por exemplo, se se introduce *costa* no campo do selector pódese elixir un ou varios dos documentos que conteñen esa palabra ou mesmo todos eles:



The screenshot shows the CORGA search interface. At the top, there is a navigation bar with the logo and menu items: CORGA, Información, Búsqueda, Guía, Frecuencias, Contacto, and Equipo. Below this, the search area is divided into several sections:

- Busca:** Includes a search bar with the text "Palab. ortográficas" and a "Corpus" dropdown set to "Etiquetado automaticamente".
- Resultado:** Shows "Tipo de resultado" as "Concordancias", "Ordenación" as "Coincidencia", "Data", and "Título", and "Tamaño de páxina" as "50".
- Sensibilidade:** Includes checkboxes for "Acentos" and "Maiúsculas", both of which are checked.
- Filtros:** A grid of dropdown menus for various filters: Orixe, Bloque, Xénero, Subtipo, Medio, Sección, Dende, Ata, Área temática, Subárea, Sexo-autor, Sexo-interlocutor, Autor, and Obra. The "Documento" filter is currently open, showing a list of documents with "Costa Norte/ZFK" selected.
- Search Results:** A search bar with the text "Palab. ortográficas" and a limit of "Cinco palabras máximo".

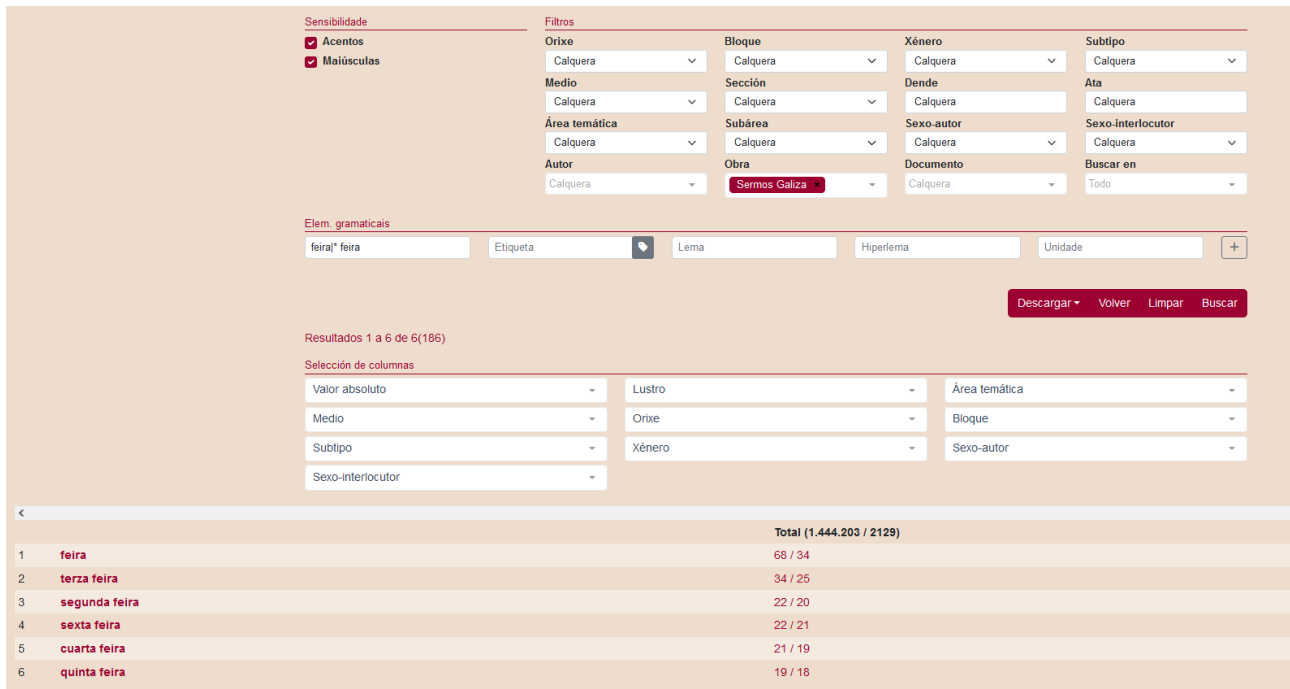
At the bottom right, there are buttons for "Volver", "Limpar", and "Buscar".

Fig. 54. Exemplo de selección dun documento específico.

A barra vertical ‘/’ separa o título da colección de relatos ou blog do título de cada un dos relatos e entradas específicos, e a barra para desprazarse permite ver o identificador completo do documento, no caso de que o seu tamaño exceda o espazo da ventá dispoñible, como pasa coa entrada do blog *Natureza Dixital. Retallos de natureza en formato dixital / Xacinto das uvas (Muscari neglectum), unha sorpresa na costa galega*, que posúe a palabra *costa* no identificador do documento pero que non podemos visualizar a non ser que empreguemos a barra de desprazamento.

6.10. Obra

O filtro *Obra* é de utilidade para agrupar nun único documento os documentos aniñados que este poida conter: noticias e relatos curtos, sobre todo, pero tamén artigos ensaísticos, por exemplo. Agora ben, no caso dos xornais e revistas a agrupación vai máis alá e supera o documento colectivo concreto, integrando todos os días do mes incorporados ao corpus ou números de distinto ano que no corpus constan dese xornal ou revista. Pensemos na posibilidade de contrastar o emprego dunha determinada palabra, por exemplo *feira*, entre A Nosa Terra (68 casos, ningún dos cales como palabra illada ou xunto con *segunda*, *terza*, *cuarta* etc. para a denominación dos días da semana) e Sermos Galiza (186 ocorrencias nas que son frecuentes usos formando parte de días da semana):



The screenshot shows the search interface for the word "feira". The filters section includes:

- Sensibilidade:** Acentos, Maiúsculas
- Filtros:**
 - Orixe: Calquera
 - Bloque: Calquera
 - Xénero: Calquera
 - Subtipo: Calquera
 - Medio: Calquera
 - Sección: Calquera
 - Dende: Calquera
 - Área temática: Calquera
 - Subárea: Calquera
 - Sexo-autor: Calquera
 - Autor: Calquera
 - Obra: Sermos Galiza
 - Documento: Calquera
 - Sexo-interlocutor: Calquera
 - Buscar en: Todo

The search criteria are: "feira" (feiral* feira), Etiqueta, Lema, Hiperfema, and Unidade. The results table shows:

		Total (1.444.203 / 2129)
1	feira	68 / 34
2	terza feira	34 / 25
3	segunda feira	22 / 20
4	sexta feira	22 / 21
5	cuarta feira	21 / 19
6	quinta feira	19 / 18

Fig. 55. Exemplo de selección dunha obra.

6.11. Clasificación textual. O parámetro *Tipo de texto*

A incorporación de transcripcións, guións e blogs ao CORGA pon de manifesto a necesidade de adaptar o sistema de consultas para darlles cabida a estes novos tipos de documentos, mais evidencia tamén a mestura que se producía ata a versión 3.0 na clasificación dos documentos entre tipoloxía de documento e tipoloxía temática. Presentóusenos, pois, a oportunidade de modificar esta situación reclasificando os documentos para discriminar neles entre tipo de documento e área temática. Deste xeito facilítaselle ao usuario non familiarizado co corpus unha clasificación dos documentos contidos nel coherente e sinxela, manexable en definitiva, e facilítaselle, así mesmo, a consulta por grandes bloques: *ficción* fronte á *non ficción*, ou *prensa* fronte ao *ensaio*, por exemplo, consultas para as que ata o de agora había que realizar varias procuras e cuxos datos só o usuario que coñecía ben o sistema era capaz de obter.

A partir da versión 3.0 a clasificación por área temática, como vimos de ver, realízase só para os textos ensaísticos e xornalísticos, mentres que todos os documentos se catalogan segundo a súa tipoloxía textual.

As variables que temos en conta para clasificar tipoloxicamente os documentos no CORGA e que, á súa vez, se poden empregar para filtrar o subcorpus no que desexan realizarse as consultas

son as seguintes: *orixe*, *bloque*, *xénero* e *subtipo*. Temos que distinguir entre o subcorpus oral e o escrito. Na parte oral só aplican dúas variables de clasificación: a *orixe* e o *subtipo*, fronte á parte escrita na que aplican todas: *orixe*, *bloque*, *xénero* e *subtipo*.

A aplicación de consulta proporciona un menú amigable para cada un dos catro parámetros anteriores, para os cales, premendo á súa vez na pestana respectiva, se listan os valores posibles. Son os seguintes.

6.9.1. Orixe

Distínguese aquí a procedencia do documento, ou sexa, se este ten orixe na escrita ou pola contra procede da transcripción dun texto oral. Se o valor escolleito é **Oral**, accédese directamente á tipoloxía na que se clasifican as transcripcións, mentres que se se opta por **Escrita** ou **Calquera** habilítase o parámetro **Bloque**.

6.9.2. Bloque

Facilítase con esta variable a distinción entre os dous grandes bloques que conforman o CORGA: **Ficción** e **Non ficción**.

6.9.3. Xénero

A clasificación por xénero permite remitir os documentos a un dos seguintes: **Dramático**, **Ensaístico**, **Narrativo** e **Xornalístico**.

6.9.4. Subtipo

Os valores aquí contidos estarán activados ou desactivados en función das escollas anteriores. Se antes non se realiza ningunha selección, estarán todas as posibilidades activadas.

Para a orixe **Oral** as opcións polo momento son as seguintes: *conferencia*, *entrevista*, *informativo*, *programa cultural*, *publicidade*, *tertulia* e *variedades*.

Para a orixe **Escrita** as posibilidades son as seguintes: *novela*, *relato curto*, *obra de teatro*, *guión*, *xornal*, *revista*, *blog*, *memoria*, *libro de texto*, *artigo científico* e *divulgación*.

En resumo, as dependencias que se establecen segundo a orixe e demais variables catalogadoras da tipoloxía textual son as seguintes:

ORIXE	BLOQUE	XÉNERO	SUBTIPO
Oral	-----	-----	Conferencia
			Entrevista
			Informativo
			Programa cultural
			Publicidade
			Tertulia
			Variedades

Táboa 2. Valores posibles para a orixe **Oral**.

ORIXE	BLOQUE	XÉNERO	SUBTIPO
Escrita	Ficción	Narrativo	Novela
			Relato curto
		Dramático	Obra de teatro
			Guión
	Non ficción	Xornalístico	Xornal
			Revista
			Blog
		Ensaístico	Memoria ¹⁰
			Libro de texto
			Artigo científico ¹¹
		Divulgación ¹²	

Táboa 3. Valores posibles e dependencias para a orixe **Escrita**.

Finalmente, debe terse en conta que cando nos resultados se di que unha determinada forma aparece en x número de documentos, a alta cifra de documentos que se observa en numerosos casos débese á concepción que no corpus posúe o documento. Así, no caso de xornais e revistas, cada noticia trátase como un documento independente. No caso dos libros, trátase como un documento independente cada prólogo ou apéndice que posúa autoría distinta da da obra xeral e, por último, no caso de coleccións de relatos ou ensaios, cada elemento da colección trátase tamén como un documento independente.

7. Notas para a interpretación dos resultados

Pode suceder que o número de ocorrencias que nos devolve o sistema nunha busca puntual por palabra ortográfica sexa moito menor ca se facemos esa mesma busca por elemento gramatical, co cal non entendemos que pode estar sucedendo ou cremos que hai un erro no funcionamento da aplicación. Percíbese esta problemática por exemplo coa consulta de *entra*.

A sensibilidade ás maiúsculas e minúsculas é a causante da diferenza, pois como elemento gramatical *entra* só pode ser forma verbal e polo tanto sempre se escribe en minúscula, pero como palabra ortográfica pode aparecer a comezo de enunciado e escribirse en maiúscula, o que fai que nas consultas por palabra ortográfica sexa pertinente a diferenza. Para que tamén fose pertinente en elemento gramatical teríamos que cubrir o campo **unidade** con *entra*, e veríamos que o número de ocorrencias totais é o mesmo.

Así pois, téñase en conta na interpretación dos resultados que manter habilitada a diferenciación entre maiúsculas e minúsculas repercute no número de resultados que devolve o sistema para a consulta sobre unha mesma forma por palabra ortográfica e por elemento gramatical.

Finalmente, pódese pensar tamén que o sistema de buscas non funciona correctamente ao ver os resultados dunha procura na que estea implicado o signo de puntuación de peche de admiración, e en menor medida o de interrogación. Se cubrimos o campo de texto coa secuencia

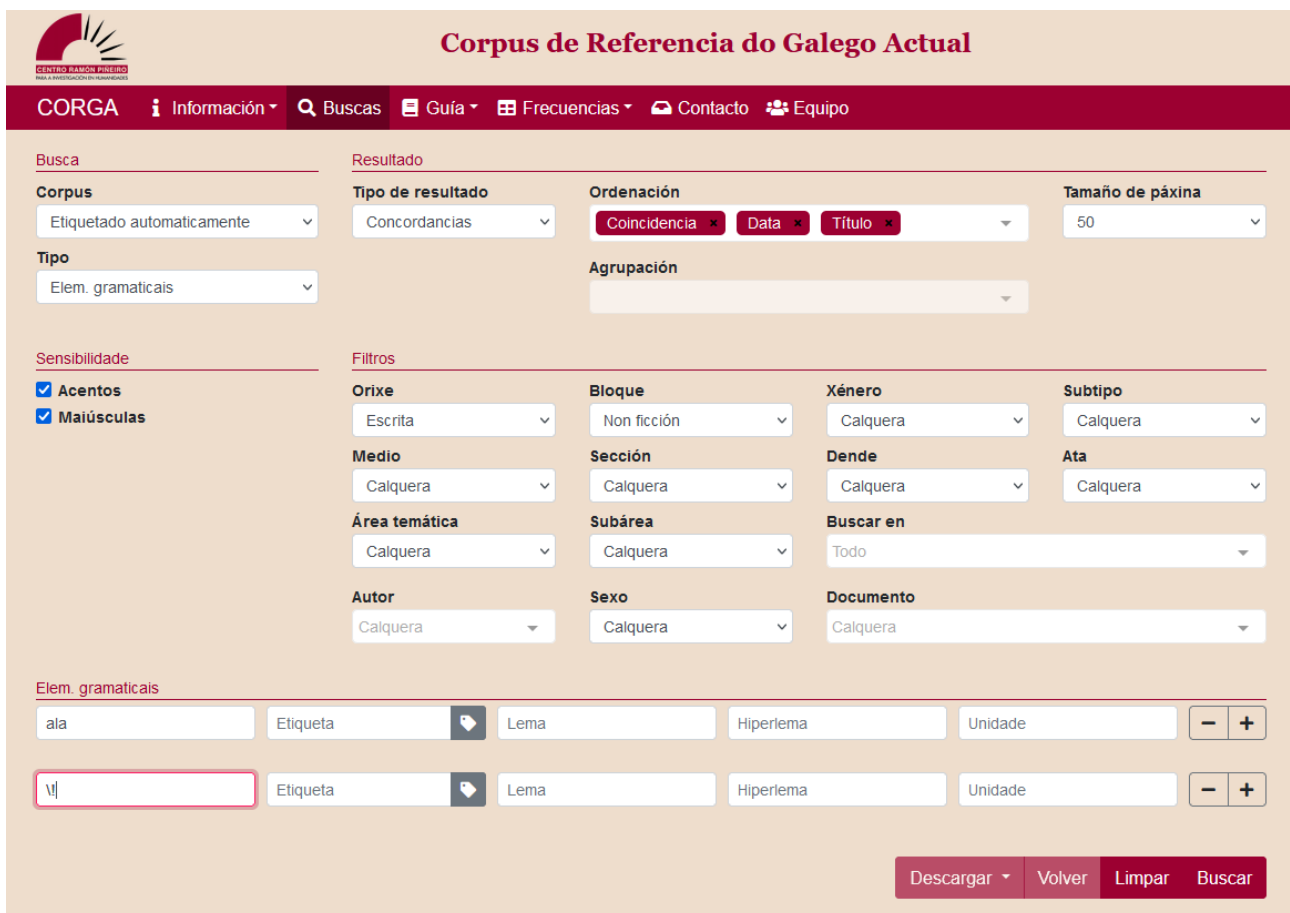
¹⁰ Inclúe as entrevistas e biografías.

¹¹ Acolle coleccións semellantes ás revistas académicas especializadas.

¹² Integra o groso do xénero ensaístico, fóra as memorias, libros de texto e artigos científicos.

ala! buscando os casos nos que esa interxección actúa como fórmula de despedida, pensaremos que o sistema toleou ao ver que nos devolve todas as ocorrencias nas que aparece *ala*, sen ter en conta que pediamos que estivese seguido do signo de admiración. Mais non é así, o sistema funciona perfectamente e devólvenos o que pedimos, porque non estamos tendo en conta que o signo de admiración de peche emprégase como operador booleano para impedir a aparición das palabras, formas, etiquetas ou lemas que precede.

Nas procuras por formas ortográficas non é posible empregar signos de puntuación, a menos que estes formen parte da forma gráfica (por exemplo nas abreviacións). Lémbrese, polo tanto, que para empregar nas buscas por elementos gramaticais ‘?’ e ‘!’ como signos de puntuación debe precedelos sen espazos entre eles a barra oblicua invertida ‘\’, acción que xa cobre o sistema cando se quere empregar algunha desas dúas etiquetas e se introducen a través do menú amigable dispoñible. Así pois, para recuperar os casos da interxección *ala* seguida do signo de puntuación de peche de admiración habería que cubrir coma na imaxe seguinte:



The screenshot shows the 'Corpus de Referencia do Galego Actual' search interface. At the top, there is a navigation bar with 'CORGA' and links for 'Información', 'Buscas', 'Guía', 'Frecuencias', 'Contacto', and 'Equipo'. Below this, the search area is divided into several sections:

- Busca:** Includes 'Corpus' (set to 'Etiquetado automaticamente'), 'Tipo' (set to 'Elem. gramaticais'), 'Sensibilidade' (with 'Acentos' and 'Maiúsculas' checked), and 'Elem. gramaticais' (with 'ala' entered in the first field).
- Resultado:** Includes 'Tipo de resultado' (set to 'Concordancias'), 'Ordenación' (with 'Coincidencia', 'Data', and 'Título' buttons), and 'Tamaño de páxina' (set to 50).
- Filtros:** A grid of filters including 'Orixe' (Escrita), 'Bloque' (Non ficción), 'Xénero' (Calquera), 'Subtipo' (Calquera), 'Medio' (Calquera), 'Sección' (Calquera), 'Dende' (Calquera), 'Ata' (Calquera), 'Área temática' (Calquera), 'Subárea' (Calquera), 'Buscar en' (Todo), 'Autor' (Calquera), 'Sexo' (Calquera), and 'Documento' (Calquera).

At the bottom, there are buttons for 'Descargar', 'Volver', 'Limpar', and 'Buscar'. The search field for 'Elem. gramaticais' is highlighted with a red box, and the text 'ala' is entered.

Fig. 56. Pantalla de captación de datos por elementos gramaticais para *ala!*