Manuel González González (ed.)

Lingua, pobo e terra

Estudos en homenaxe a Xesús Ferro Ruibal

Xunta de Galicia: Centro Ramón Piñeiro para a investigación en humanidades
Santiago de Compostela - 2016

Lingua, pobo e terra: estudos en homenaxe a Xesús Ferro Ruibal / edición a cargo de Manuel González González – Santiago de Compostela: Xunta de Galicia. Centro Ramón Piñeiro para a investigación en humanidades, 2016.

544 p.; 24 cm. Homenaxes – Centro Ramón Piñeiro para a investigación en humanidades

D.L. C 1602-2016 ISBN: 978-84-453-5236-6

1. Lingua galega 2. Lexicografía galega 3. Fraseoloxía 4. Historia da lingua galega 5. Onomástica 6. Dialectoloxía 7. Corpus I. Ferro Ruibal, Xesús II. González González, Manuel III. Xunta de Galicia: Centro Ramón Piñeiro para a investigación en humanidades.

Edita: Xunta de Galicia. Centro Ramón Piñeiro para a investigación en humanidades

O noso agradecemento a María Carbajo pola axuda prestada na uniformización dos textos

Maquetación e impresión:
Grafisant, S.L.

ISBN:
978-84-453-5236-6

Depósito legal:
C 1602-2016

O Corpus de Referencia do Galego Actual (CORGA): estado actual e perspectivas

Guillermo Rojo
Marisol López Martínez
Eva Mª Domínguez Noya
Fco. Mario Barcala
Centro Ramón Piñeiro para a Investigación en Humanidades

1. Introdución

O *Corpus de Referencia do Galego Actual* (CORGA) é, sen dúbida, un dos proxectos máis representativos do labor desenvolvido polo *Centro Ramón Piñeiro para a Investigación en Humanidades* (CRPIH) nos seus vintetrés anos de existencia, período case exactamente coincidente coa presenza e participación activa nel de Xesús Ferro Ruibal, que formou parte do Centro dende os seus comezos. En efecto, cando alá polo 1993 Constantino García estableceu o primeiro deseño das áreas de traballo do CRPIH, situou no seu núcleo a integración harmónica e sinérxica da informática e as humanidades. A creación dun corpus textual amplo, equilibrado e representativo do galego contemporáneo foi unha consecuencia absolutamente natural das funcións atribuídas ao CRPIH no seu decreto fundacional.

Por aqueles anos, a lingüística de corpus adquirira xa un altísimo nivel de logros baseados en tres aspectos distintos. En primeiro lugar, a demostración de que fronte á marxinación que esta aproximación sufrira e sufría aínda nos Estados Unidos, os corpus creados en diversos países europeos (principalmente o Reino Unido e os países nórdicos) resultaban ser un medio de gran rendibilidade mesmo para os estudos de corte máis teórico e, por suposto, os aplicados. En segundo termo, o desenvolvemento de linguaxes de codificación (SGML daquela) que permitían marcar un texto de xeito que reflectise fielmente a súa estrutura conceptual e organizativa e, polo tanto, permitise a creación dinámica de subcorpus virtuais e o que agora chamamos recuperación selectiva da información. Por fin, o desenvolvemento da lingüística computacional, grazas á cal os textos podían ser anotados morfosintacticamente e lematizados de xeito automático, co que a recuperación de información podía acadar independencia respecto da forma ortográfica e, polo tanto, adquirir a posibilidade de facer buscas de carácter abstracto, do

Manuel González González (ed.)

Lingua, pobo e terra

Estudos en homenaxe a Xesús Ferro Ruibal

Xunta de Galicia: Centro Ramón Piñeiro para a investigación en humanidades
Santiago de Compostela - 2016

Lingua, pobo e terra: estudos en homenaxe a Xesús Ferro Ruibal / edición a cargo de Manuel González González – Santiago de Compostela: Xunta de Galicia. Centro Ramón Piñeiro para a investigación en humanidades, 2016.

544 p.; 24 cm. Homenaxes – Centro Ramón Piñeiro para a investigación en humanidades

D.L. C 1602-2016 ISBN: 978-84-453-5236-6

1. Lingua galega 2. Lexicografía galega 3. Fraseoloxía 4. Historia da lingua galega 5. Onomástica 6. Dialectoloxía 7. Corpus I. Ferro Ruibal, Xesús II. González González, Manuel III. Xunta de Galicia: Centro Ramón Piñeiro para a investigación en humanidades.

Edita: Xunta de Galicia. Centro Ramón Piñeiro para a investigación en humanidades

O noso agradecemento a María Carbajo pola axuda prestada na uniformización dos textos

Maquetación e impresión:
Grafisant, S.L.

ISBN:
978-84-453-5236-6

Depósito legal:
C 1602-2016

estilo das habituais nas investigacións lingüísticas (por lemas ou categorías gramaticais, por exemplo). E estes tres aspectos eran facilitados e integrados polo desenvolvemento das computadoras, que permitían xa o almacenamento de grandes masas de datos, o seu procesamento a alta velocidade e a interconexión mundial a través de Internet.

Todos estes compoñentes actuaban xa no momento no que o CORGA foi concibido e materializáranse no *British National Corpus* (BNC), rematado en 1994, e integrado pola portentosa cifra (para a época) de cen millóns de formas procedentes de textos escritos e orais do inglés británico contemporáneo. O BNC foi, á distancia esperable polos medios económicos dispoñibles e os condicionamentos sociais das linguas respectivas, o modelo sobre o que se deseñou o CORGA.

Na situación actual, os que chamamos corpus de referencia —é dicir, os que están preparados para proporcionaren información xeral acerca dunha lingua— están situados na zona media entre dúas tendencias distintas e, en certo modo, opostas. Por unha banda, os corpus de tamaño pequeno (digamos cinco millóns do formas ou menos), que poden ter un coidado exquisito na selección, edición e codificación dos textos que os compoñen, en moitos casos traballados expresamente para a súa integración no corpus correspondente. A especialización nun determinado tipo de texto permite mesmo incluír materiais complementarios, como, por exemplo, os manuscritos, a edición paleográfica, imaxes asociadas, etc. A *base de datos da Lírica Profana Galego-Portuguesa*, elaborada tamén no CRPIH, ten moitas das posibilidades apuntadas con relación a este corpus textual reducido e aceptablemente homoxéneo. Son os corpus que, en iluminadora caracterización de Mair (2006), resultan 'small and tidy'. Como é lóxico, nestas características reside tamén a súa limitación fundamental: están constituídos por un conxunto pequeno de textos que, amais, son do mesmo tipo ou de tipos moi estreitamente relacionados.

No extremo contrario están os corpus que Mair caracteriza como «big and messy». Son os formados a partir da reunión de materiais xa preexistentes en Internet. De forma altamente automatizada, é relativamente doado construír corpus que acadan, en non moito tempo e sen necesidade de inversións económicas, milleiros de millóns de formas. Resultan moi útiles para a análise de fenómenos ou elementos de frecuencia baixa, pouco ou nada representados nos corpus de referencia. A actuación de programas de

anotación e lematización automáticas fai que as aplicacións de busca asociadas poidan tamén traballar con elementos abstractos, como, por exemplo, as categorías gramaticais. A súa limitación básica reside no seu carácter 'oportunista' e, sobre todo, na imposibilidade de facer recuperación selectiva de información en sentido estricto, dado que o seu tamaño obriga a reducir a codificación á que cada texto integrado ten na súa forma orixinal, moi variada por outra parte.

O CORGA está, como todos os corpus de referencia, entre estes dous extremos. Pretende acadar, como se mostra nos apartados seguintes, un tamaño suficiente para facer posible a análise dos elementos e fenómenos que se dan nunha lingua determinada, en textos dos máis diversos tipos. Polo tanto, teñen como característica fundamental a existencia dun deseño previo que indica, por exemplo, a porcentaxe de textos de cada medio, tipo ou período que van entrar na súa composición. Da existencia do deseño deriva a riqueza da codificación –que ten que reflectir os valores de cada texto nos diferentes parámetros que entran na súa configuración– e a posibilidade de recuperación selectiva (exemplos dun certo elemento en textos de tal tipo, tal ano, etc. ou ben en zonas específicas do texto, como o titular dunha noticia ou o prólogo dunha novela).

A pregunta acerca de se compensa construír un corpus de referencia, con unicamente ducias de millóns de formas cando é posible facer consultas directamente en Internet, con buscadores comerciais, sobre centos ou milleiros de millóns de palabras queda contestada nos parágrafos anteriores. O que se persegue habitualmente na investigación lingüística non é tanto a frecuencia total dun determinado fenómeno coma a diferente frecuencia con que se dá en textos de distintas características (tipo, medio, período, zona xeográfica, etc.). Evidentemente, este trazo esixe a codificación previa, tarefa custosa en tempo e que, polo tanto, require importantes recursos económicos. Por outra parte, en Internet hai moitos textos, pero, por diferentes razóns, non hai moitos de certos tipos. Se o CORGA pode conter, por exemplo, obras literarias publicadas recentemente, é porque autores e grupos editoriais aceptaron a súa inclusión precisamente pola garantía de que os usuarios do corpus poden recuperar liñas de concordancias, pero nunca a obra completa nin fragmentos representativos dela. En definitiva, os corpus de referencia en xeral e o CORGA en particular constitúen recursos insubstituíbles na investigación lingüística do galego, como intentamos poñer de manifesto na exposición das súas características que desenvolvemos nos apartados seguintes.

2. O Corpus de Referencia do Galego Actual (CORGA)

2.1. Parámetros de configuración

O Corpus de Referencia do Galego Actual (CORGA) é un corpus documental aberto integrado por distintos tipos de textos representativos da lingua galega actual escrita —xornais, semanarios, revistas, ensaios, e textos de ficción (novela, relato curto e teatro)—, codificados no estándar XML (*eXtensible Markup Language*) e que abrangue temporalmente dende o ano 1975 ata a actualidade. Neste momento dispone dunha versión en liña con 31,9 millóns de formas ortográficas á que se pode acceder a través de http://corpus.cirp.es/corga.

Os textos que forman parte do CORGA foron seleccionados de acordo con tres criterios de clasificación, independentes entre si: *tipo de texto*, *data* e *área temática*. A distribución realizouse por lustros, procurando darlle maior representatividade aos períodos máis recentes[1]:

Tipo de texto	Data
Xornal	1975-1979
Revista	1980-1984
Ensaio	1985-1989
Novela	1990-1994
Relato curto	1995-1999
Teatro	2000-2004
	2005-2009
	2010-2014

Táboa 1. Valores dos parámetros de clasificación Tipo de texto e Data.

1 Non se confunda distribución de documentos por lustro para a selección de documentos que se introducen no CORGA con *ano* parámetro no sistema de recuperación de información, onde o período cronolóxico selecciónao o usuario segundo os seus intereses: todo o corpus, un ano concreto ou un rango específico de anos, loxicamente dentro do período temporal que abrangue o corpus.

	ÁREAS TEMÁTICAS					
	Economía e política	**Cultura e artes**	**Ciencias sociais**	**Ciencias e tecnoloxía**	**Outros**	**Ficción**
SUBÁREAS TEMÁTICAS	Política	Audiovisuais e espectáculo	Lingua	Sanidade	Deportes	Novela
	Desenvolvemento e infraestruturas	Medios de comunicación	Literatura	Bioloxía, botánica, ecoloxía, zooloxía e paleontoloxía	Turismo	Relato curto
	Emprego, traballo, industria	Artes gráficas e plásticas	Relixión	Tecnoloxía e industria	Afeccións e asuntos domésticos	Teatro
	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 2. Valores das áreas e subáreas temáticas nas que se clasifican os textos do CORGA.

2.2. Datos da versión actual

Ata o momento incorporáronse ao CORGA 610 xornais (exemplares completos d'*A Nosa Terra, A Peneira, De Luns a Venres, Diario Oficial de Galicia, Galicia Hoxe, O Correo Galego, O Xornal de Galicia* e *Sermos Galiza* e noticias soltas de *La Voz de Galicia* cando non existía aínda ningún xornal en galego), 126 revistas (entre elas números de *Código Cero, Cerna, Consumer Eroski, Díxitos, Entregas de Comunicación Cultural, Feiraco, Galicia Internacional, Man Común, Petroglifo, Revista Galega de Economía, Teatro do Noroeste, Teima* ou *Tempos Novos*) e 509 libros (168 novelas, 140 ensaios, 123 coleccións de relato e 78 obras de teatro).

No que se refire ao tamaño, a última versión do CORGA, a 1.7, consta de 31.977.962 palabras (sen incluír as cifras). Hai, non obstante, un subconxunto de 382.055 formas gráficas sobre as que non se poden facer buscas, pois son unidades que aparecen en fragmentos textuais que están nunha lingua distinta do galego ou aparecen en notas de tipo bibliográfico, resultando un total de 31.595.907 formas ortográficas sobre as que si se poden realizar consultas, das cales 430.376 son formas diferentes.

A amplitude de documentos dos que consta o CORGA así como os criterios empregados na súa selección permítennos considerar este corpus representativo do uso lingüístico do galego actual.

A distribución das frecuencias das formas ortográficas segundo os parámetros de selección dos documentos son as seguintes:

Distribucións por medio		
Medio	*Frecuencia*	*Porcentaxe*
Xornal	11.030.959	34.91 %
Revista	2.715.420	8.59 %
Libro	17.849.528	56.49 %
	Ficción: 11.697.773	37.02 %
	Non ficción: 6.151.755	19.47 %

Táboa 3. Distribución das frecuencias por medio.

Distribucións por período		
Período	*Frecuencia*	*Porcentaxe*
1975-1979	751.728	2.38 %
1980-1984	1.305.122	4.13 %
1985-1989	1.581.247	5.00 %
1990-1994	4.208.337	13.32 %
1995-1999	7.743.733	24.51 %
2000-2004	5.928.485	18.76 %
2005-2009	7.631.055	24.15 %
2010-2014	2.446.200	7.74 %[2]

Distribucións por área temática principal		
Área temática	*Frecuencia*	*Porcentaxe*
Economía e política	7683404	24.32 %
Cultura e artes	2460913	7.79 %
Ciencias sociais	4291045	13.58 %
Ciencia e tecnoloxía	2368553	7.50 %
Outros	3094219	9.79 %
Ficción	11697773	37.02 %

Táboa 4. Distribución das frecuencias por período (esquerda) e por área temática principal (dereita).

2 Estase traballando aínda na incorporación de documentos pertencentes ao último período, o que explica esta baixa porcentaxe.

2.3. A codificación e estruturación dos documentos

Os documentos que se incorporan ao CORGA codifícanse segundo o estándar XML (*eXtensible Markup Language*) co fin de incrementar as posibilidades de recuperación de información e, ademais, garantir a permanencia no tempo. Esta codificación implica un deseño e estruturación que dá conta da disposición interna característica de cada un dos grandes tipos de textos (xornal, teatro, ensaio, etc.). Así, exemplificando cun texto xornalístico, esta estruturación permite considerar un xornal un único documento que está organizado en noticias, distribuídas en *seccións*, as cales, á súa vez, conteñen obrigatoriamente un *corpo* e opcionalmente un *titular*, *resumo* e/ou *pé de foto*. A maiores, cada un destes elementos está constituído por parágrafos (texto comprendido entre dous puntos e á parte), e estes son segmentados en oracións (secuencia textual separada do resto do texto por un signo forte de puntuación). Naturalmente, o xornal posúe ademais unha cabeceira complexa na que se recollen os datos bibliográficos e unha cabeceira específica por noticia onde se inclúen os datos relativos a autor, sección e áreas temáticas que correspondan.

Esta disposición detallada habilita a posibilidade de, no sistema de recuperación de información, unha vez indexados os documentos, realizar consultas sobre a totalidade do documento (*noticia*, seguindo co exemplo) ou sobre unha unidade estrutural concreta (para o tipo de documento xornal: *titular, resumo, pé de foto* ou *corpo*).

A maiores das unidades estruturais xa citadas que dan conta da organización interna das noticias que poden conformar xornais ou revistas (*corpo, titular, resumo* e *pé de foto*), tivéronse en conta na estruturación interna outras que adoitan aparecer no tipo de documento *libro*. Comparten, así mesmo, outra característica: todas son opcionais, pois o único elemento constitutivo obrigatorio é *corpo*. Vexámolas:

> **Prólogo**: unidade estrutural do contido do documento que precede o *corpo* e baixo a que se engloban as distintas denominacións: *presentación, limiar, introdución, preámbulo, prefacio,* etc. Cando o prólogo é dun autor diferente do da obra xeral incorpora unha cabeceira coas características deste: autor, título se o ten, áreas temáticas, sistema ortográfico e/ou comentarios do lingüista revisor.
>
> **Apéndice**: unidade estrutural do contido do documento que segue o *corpo*. En realidade, se hai *apéndice* este é o último elemento constituti-

vo do documento. Baixo el englóbanse *epílogo, agradecementos, glosario, táboa de feitos histórica, glosario, vocabulario,* etc.

Dedicatoria: non necesita aclaración.

Cita: non necesita aclaración.

Encabezamento: como o seu nome indica, é o título co que se encabezan capítulos, apartados ou epígrafes. A semellanza do *titular*, presenta para o estudo da lingua unha sintaxe especial.

Nota: unidade na que se recolle a explicación, advertencia ou comentario de calquera tipo que vai fóra do texto, ben en nota ao pé ben en notas finais. Este tipo de nota sempre se referencia no texto, normalmente mediante numeración en superíndice.

Nota tipo 2: unidade na que se recolle a explicación, advertencia ou comentario de calquera tipo situada fóra do texto pero que non está referenciada.

Todos os elementos estruturais descompóñense en parágrafos, e estes, á súa vez, en oracións, sendo esta última a unidade sobre a que se realizan as buscas no sistema de consultas e, loxicamente, tamén a que se recupera, aínda que logo se poida ampliar o contexto a dúas oracións anteriores e dúas posteriores.

Cómpre subliñar que os textos introducidos no CORGA non conservan o formato: cursivas, grosas ou emprego de versais como elemento marcador desaparecen cando se prepara o documento para a súa inclusión no corpus. Tampouco se mantén a paxinación. Débese ter en conta que, unha vez que un texto se converte en documento electrónico, deixa de ter importancia a localización por páxina do orixinal, cuxo mantemento suporía cortar unidades, pois as localizacións fanse doutros xeitos. Non se cita xa a ocorrencia dunha forma dada na páxina *n* de tal obra, senón que a referencia se fai pola aparición da forma *x* na obra *z* que aparece no corpus, no noso caso, CORGA.

A codificación que se lles aplica aos textos que se introducen no CORGA inclúe, á parte da estruturación nas unidades anteriores segundo corresponda, unha serie de etiquetas que achegan información de diverso tipo. As principais son as seguintes:

Distinto: en función do valor concreto que o complete (*outra lingua, outro período, non normativo* e *descoñecido*), indica respectivamente que

o texto que engloba pertence a unha lingua diferente do galego, a un estadio anterior da nosa lingua (anterior a 1955), é texto reintegracionista ou correspóndese cunha lingua inventada (falar de cans, corvos, etc.). Márcanse sempre coa etiqueta «distinto» os parágrafos e oracións que responden ás características anteriores. Agora ben, cando se trata de palabras soltas non se marcan a non ser que apareza a continuación a tradución ou explicitamente se indique que é unha unidade pertencente a outra lingua. O obxectivo é impedir que o texto que non pertence ao galego actual produza ruído no cómputo de formas que constitúen o corpus ou na recuperación da información. O texto englobado nesta etiqueta non pode consultarse aínda que si verse contextualmente, destacado no sistema de consultas en cor amarela.

Poema. Por unha banda cómpre integrar os poemas ou fragmentos que aparecen nos textos que se incorporan ao corpus; por outra, o significado e sintaxe das formas que aparecen en poemas non corresponden a un uso habitual e representativo dunha lingua. É por isto que se crean as etiquetas *inicio_poema, fin_poema* e *novo_verso*.

Táboa. Cando unha táboa non é representable, tendo en conta as posibilidades de estruturación que definimos, indícase con esta etiqueta que alí onde aparece, no orixinal, hai unha táboa. A maiores, *táboa* é tamén un atributo do *parágrafo* que empregamos cando existe unha táboa representable, en xeral que consta dunha columna e que contén vocabulario inusual. Dado que maioritariamente as oracións en que se desagregan as táboas non son oracións ao uso, senón simplemente palabras ou frases, a etiqueta *táboa* axúdalle ao usuario a entender os resultados.

Fórmula. Emprégase esta etiqueta cando existe unha fórmula para cuxa representación teriamos que botar man de caracteres especiais.

Interlocutores en entrevistas ou personaxes dunha obra de teatro. O obxectivo é impedir que desvirtúen as frecuencias aparecendo, poñamos por caso, 200 Avelino, 200 Pousa e 200 Antelo.

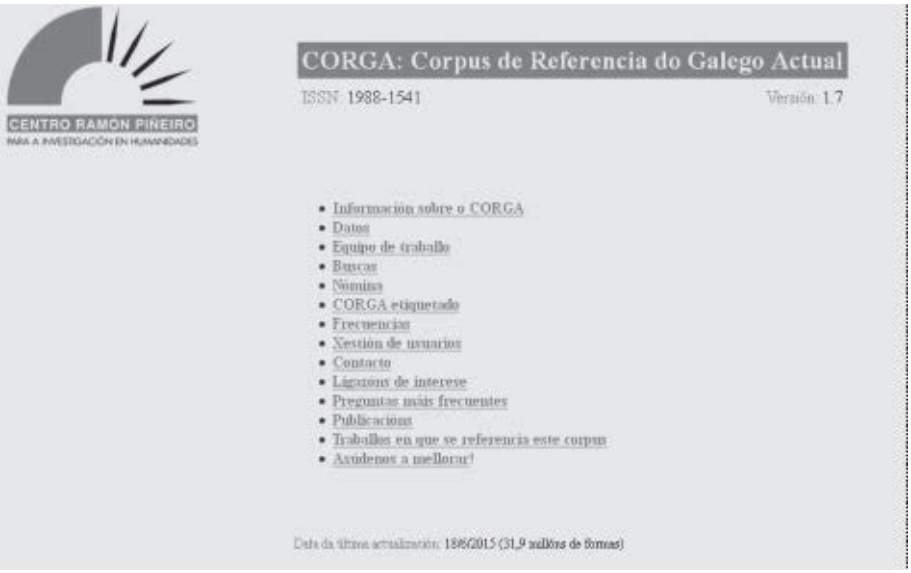
Referencia bibliográfica. É unha etiqueta que se lle asigna á marca estrutural *nota* cando, como o seu nome indica, se trata dunha simple referencia bibliográfica. Polo xeral son oracións que só conteñen títulos en linguas diferentes do galego, autores e topónimos, motivo polo que

se considera que as formas que aí aparecen non teñen interese para a recuperación de información e non son indexadas.

Acoutación. Reservado exclusivamente para as obras de teatro, *acoutación* é un valor que se lle atribúe aos parágrafos e/ou oracións que conteñen as partes non dialogadas do texto dramático onde se atopan as indicacións escénicas (precisións sobre o lugar, xestos, apariencia física, vestiario etc.), ben orientadas á lectura, ben orientadas á posta en escena.

2.4. A aplicación de consulta

A aplicación organízase en varias seccións ás que se accede premendo en cada unha das denominacións:



Información sobre o CORGA e *Equipo de traballo* dan conta, respectivamente, do tipo de corpus que é o CORGA, os criterios empregados na selección dos documentos que se incorporan e mais o equipo de traballo autor do corpus.

Datos, pola súa banda, especifica a distribución de formas segundo cada un dos parámetros de selección, a listaxe de documentos incorporados, o número de formas da versión actual e información varia sobre a codificación dos textos.

Publicacións enumera as publicacións resultantes do proxecto, mentres que en *Traballos en que se referencia este corpus*, como indica o seu nome, relació-

nanse os traballos que empregaron, e así o recoñecen, o CORGA como fonte para o seu estudo.

Ligazóns de interese remite a outros corpus existentes para as linguas peninsulares, en tanto que *CORGA etiquetado* enlaza coa aplicación de consulta do CORGA etiquetado e lematizado automaticamente que contou con supervisión lingüística (Domínguez, 2013), e do que falaremos na epígrafe seguinte.

Xestión de usuarios xestiona o rexistro como usuario no sistema para poder realizar consultas en calquera das aplicacións.

Contacto proporciona un enderezo de correo electrónico para interactuar cos responsables do CORGA e expoñer problemas relacionados coas consultas, resolver dúbidas ou expoñer propostas en relación co corpus. Na mesma liña está *Axúdenos a mellorar!*, sección individualizada para suscitar a participación dos usuarios co envío de mensaxes comunicando erros ou fallos de funcionamento.

Finalmente, as dúas últimas seccións con información estática son *Preguntas máis frecuentes*, onde se atopa resposta ás cuestións máis usuais en relación coa consulta do corpus, e *Frecuencias*, onde se poden consultar e/ou descargar as mil formas máis frecuentes do CORGA, as cinco mil máis usuais ou a listaxe completa de frecuencias do corpus (Rojo *et al.*, 2015).

En *Buscas* accédese ao sistema de consultas. Nel pode buscarse unha palabra completa, parte dunha palabra ou varias palabras ou partes, contiguas ou non. Segundo o tipo que se escolla consultarase:

Todas as palabras: É a opción non marcada para a consulta dunha palabra, parte dunha palabra ou varias palabras que non teñen por que ser consecutivas. Nos resultados devolve as oracións que conteñan a palabra ou palabras do apartado *Palabras* en calquera posición, non necesariamente contiguas.

Calquera palabra: É a opción específica para obter ocorrencias nas que, de entre varias formas introducidas no campo de consulta, basta con que apareza unha delas. O resultado devolve as oracións que conteñan calquera das palabras que se escriba.

Frase exacta: É a opción a escoller cando se desexa recuperar unha construción con varias palabras contiguas. O resultado devolve as oracións que conteñen a frase obxecto de consulta.

Frase exacta (casos): Idéntica consulta á anterior, pero a diferenza dela, devolve o número de casos que se atoparon e non o número de oracións. As buscas que se realizan con esta opción permiten obter tamén nas estatísticas os datos sobre os casos e non sobre as oracións que cumpren o criterio de busca, coma nos demais tipos de consulta.

Booleana: É a opción a escoller cando se precisa realizar buscas complexas. En función de que a restrición sexa de tipo negativo (NOT), aditivo (AND), optativo (OR) ou aproximativo (NEAR) formularase a consulta. Así:

> AND: atopa as oracións nas que consten as dúas palabras: *accidente* AND *avión*
>
> OR: atopa as oracións nas que apareza calquera das palabras: *accidente* OR *avión*
>
> NOT: atopa as oracións nas que apareza a palabra1 pero non estea a palabra2: *accidente* NOT *avión*
>
> NEAR: atopa as oracións nas que as palabras están a unha distancia inferior a *x*, que debe especificarse: NEAR ((*non só, senón tamén*), 10)

Cando existe a posibilidade de que a consulta devolva resultados nos que a orde das palabras é variable e non interesa esa opción, pódese indicar con TRUE que as palabras que deben estar próximas nunha distancia inferior ou igual á especificada deben aparecer na orde que se establece na escrita. Por exemplo, quérese comprobar se coas formas *coche* e *liña* pode aparecer algunha construción máis, á parte da de *coche de liña*, porén queremos evitar que nos resultados poidan aparecer casos referidos ao deseño ou *liña do coche*. Para iso recorremos a TRUE e indicamos: NEAR ((*coche*, *liña*), 3, TRUE)

Á riqueza dos tipos de busca posibles debe engadirse, así mesmo, o emprego en calquera deles de dous comodíns cos cales é posible substituír caracteres:

> ? substitúe un carácter: *de?de* devolve tanto os casos de *desde* coma *dende*.
>
> * substitúe cero, un ou varios caracteres. Por exemplo, para ver que adverbios en *-mente* poden aparecer na construción booleana que viamos para NEAR cómpre indicar: NEAR ((*non *mente, senón tamén*), 10) e así nos resultados aparecerán as construcións con *somente, soamente, unicamente, únicamente* e *simplemente*.

Naturalmente, nas consultas poden cruzarse *tipo de busca* e comodíns, e incluso nas de tipo *booleano* é posible tamén combinar os operadores mediante parénteses, co que, no afán de evitar ruído e afinar os resultados que interesan, poden realizarse consultas extremadamente complexas.

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.

Por outro lado, o sistema permite a recuperación de datos da totalidade do corpus ou ben do subcorpus virtual creado directamente por quen fai a consulta en función dos diferentes parámetros utilizados na configuración do CORGA e da estruturación dos seus textos:

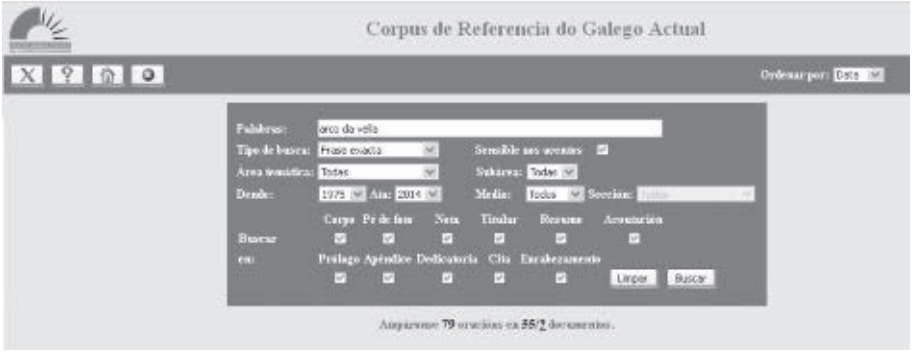
Área temática

Subárea temática

Tipo de documento

Período cronolóxico

Unidade estrutural



É posible, ademais, ordenar os resultados segundo a *data*, o *medio* ou o *tema*, e mesmo gardalos nun arquivo xml para tratalos posteriormente cando o número de resultados non permite a toma de datos inmediata.

Unha vez que se realiza a consulta pódese acceder directamente a ver os exemplos nos que se localiza a expresión de busca ou ben acceder ás esta-

tísticas de oracións e documentos clasificadas, respectivamente, por *medio*, *área temática* e *lustro*. Por exemplo, estas son as estatísticas para a palabra *beizos*:

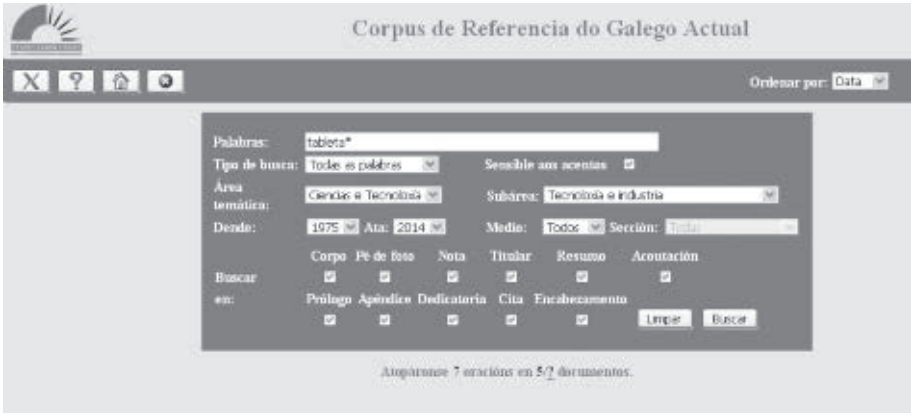
Oracións					
Estatísticas por medio		Estatísticas por área temática		Estatísticas por lustro	
Xornais	81	Economía e Política	58	1975-1979	44
Revistas	21	Cultura e Artes	18	1980-1984	111
Libros	1403	Ciencias Sociais	91	1985-1989	149
		Ciencias e Tecnoloxía	27	1990-1994	231
		Ficción	1353	1995-1999	321
		Outros	59	2000-2004	304
				2005-2009	300
				2010-2014	45
				Total: 1505	

Documentos					
Estatísticas por medio		Estatísticas por área temática		Estatísticas por lustro	
Xornais	66	Economía e Política	36	1975-1979	21
Revistas	20	Cultura e Artes	15	1980-1984	24
Libros	405	Ciencias Sociais	70	1985-1989	20
		Ciencias e Tecnoloxía	18	1990-1994	93
		Ficción	379	1995-1999	107
		Outros	41	2000-2004	79
				2005-2009	119
				2010-2014	28
				Total: 491	

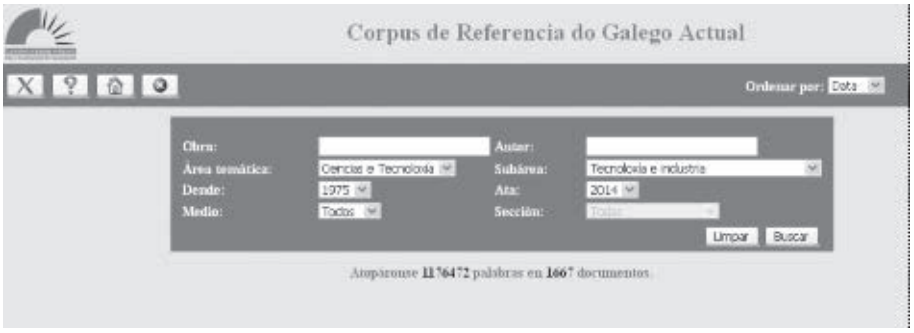
Outro dos valores engadidos do CORGA é o sistema de consulta da nómina de autores e obras que figuran no corpus. Trátase dun sistema de consulta que permite buscar que obras ou autores están no corpus, combinando esa información cos valores pertencentes aos parámetros de *ano*, *medio* e *área temática*. Accédese a el, ben directamente dende a páxina principal do CORGA ou ben enlazando a través do sistema de buscas premendo no interrogante que aparece no resultado inicial dunha consulta no número de documentos (ou cando a consulta é do tipo *Frase exacta (casos)* xunto co número de casos). Deste xeito, dependendo dos filtros que se especifiquen nas buscas, sábese que tamaño ten o subcorpus sobre o que se traballa.

Por exemplo, se escribimos *tableta** no campo de texto *Palabras* 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 tabletas* aparecen nun total de 7 oracións en 5 documentos diferentes:



E se se preme no interrogante obtemos o subcorpus sobre o que se realiza a busca.



Isto permite saber que na área e subárea especificadas o CORGA conta con 1.176.472 palabras distribuídas en 1667 documentos diferentes. O alto número de documentos que se observa para esta área concreta 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.

2.5. Liñas de traballo actuais

Co fin de situar o galego no nivel das demais linguas peninsulares, cómpre dar un salto cualitativo no propio desenvolvemento do CORGA Para iso precísase atender dúas frontes:

Por un lado, cómpre continuar incorporando novos documentos ao corpus correspondentes a textos pertencentes aos xéneros xa definidos —xornalístico, ensaístico e ficción—, para progresivamente aumentar o seu volume textual e conseguir que siga sendo representativo do uso do galego actual.

Porén, as dificultades cada vez maiores para obter documentos orixinais en galego de todos os xéneros e unha presenza vizosa do galego en Internet aconsellan dar cabida no CORGA a textos procedentes deste novo medio. A presenza do galego en Internet esixe que se documente este no corpus, tanto en relación co vocabulario procedente das novas tecnoloxías como co uso da lingua nun soporte amplamente difundido e que non se recolle aínda en ningún outro corpus do galego. Así, na actualidade trabállase no procesamento de blogs, o que supón todo un reto para a súa incorporación ao corpus por seren documentos dinámicos e posuíren unha estrutura constitutiva diferente das tratadas ata o de agora. Máis adiante, talvez, se incorporen chíos (*tweets*) ou outros semellantes.

Por outra banda, unha das carencias do CORGA nestes momentos é a súa nula representatividade para o rexistro oral. Para subsanar esta pexa, iniciouse a incorporación de textos procedentes da oralidade nunha dobre dirección:

1. Guións de series da TVG. Está previsto incorporar nos próximos anos guións, entre outras, das series *Pazo de familia, Terra de Miranda, Padre Casares, Matalobos, As leis de Celavella* ou *Casa Manola*, así como algún guión correspondente aos divulgativos *De quen vés sendo?, Náufragos* ou *Somos ben curiosos* ou algún, finalmente, do programa de humor *Era visto!* O seu procesamento é semellante ao das obras de teatro, coa particularidade de que serán textos escritos nos que predominen as características da oralidade non espontánea.

2. Paralelamente trabállase tamén na transcrición ortográfica de programas de radio gravados na década dos 90: *Camiño de volta, Pensando en ti, informativos* etc. A peculiaridade das transcricións radica en que se produce aliñando texto e voz, co que, cando os textos orais

se indexen e estea dispoñible en liña, o usuario terá non só a posibilidade de realizar buscas en textos orais, senón tamén a posibilidade de escoitar a parte do audio no que se localizan os resultados obtidos. Isto supón un proceso totalmente novidoso e converterá o CORGA nun recurso aínda máis valioso, ademais de fortalecer a presenza do galego na vangarda das tecnoloxías asociadas ao procesamento da linguaxe natural.

3. O Corpus de Referencia do Galego Actual etiquetado (CORGAetq)

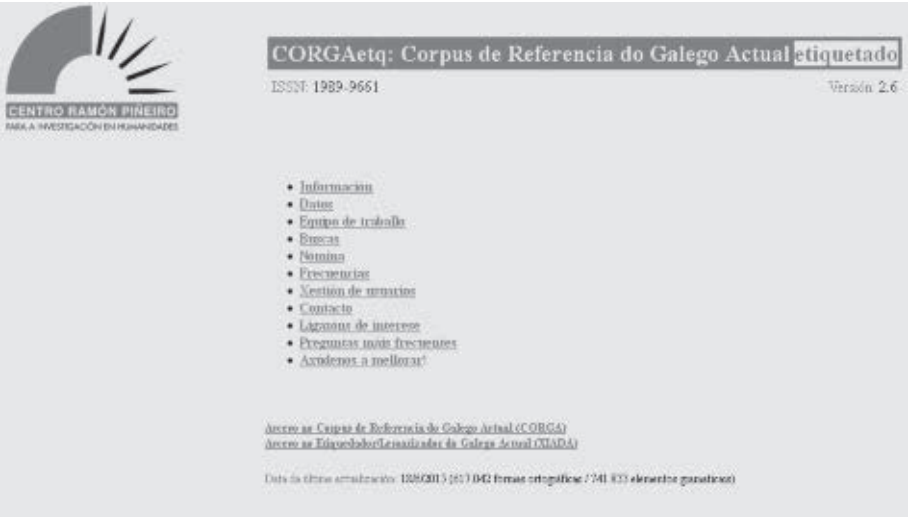
Conscientes das limitacións que impoñen as consultas por forma ortográfica e das facilidades para a recuperación de información que supón contar con etiquetas e lemas, paralelamente á construción do CORGA, o Centro Ramón Piñeiro para a Investigación en Humanidades e o grupo COLE das universidades da Coruña e Vigo estivo traballando no desenvolvemento do Etiquetador/Lematizador do Galego Actual (XIADA), destinado a etiquetar automaticamente os documentos do CORGA[3]. Unha versión deste etiquetador, posuidor dunha alta taxa de acerto (Domínguez *et al,*. 2009), está dispoñible en liña na web http://corpus.cirp.es/xiada en modo demostración, e pode utilizarse para etiquetar o texto que se lle proporcione, sempre que non exceda de 4 oracións de ata, como máximo, 512 caracteres.

Este etiquetador aplicouse sobre un subconxunto do CORGA formado por 617.042 formas ortográficas (correspondentes a 741.833 elementos gramaticais). O subcorpus foi etiquetado automaticamente e revisado á man por unha lingüista. A razón de ser deste subcorpus é servir de corpus de adestramento do xénero xornalístico e do de ficción para o etiquetador mais, dado o seu tamaño e a minuciosidade da etiquetaxe practicada nel, considérase que é de utilidade para o estudo de aspectos gramaticais, fundamentalmente. É por isto que o corpus de adestramento se puxo a disposición pública, co nome de Corpus de Referencia do Galego Actual etiquetado (CORGAetq), baixo un sistema de buscas que permite realizar consultas non só empregando formas ortográficas senón tamén lemas e etiquetas morfosintácticas. Este novo sistema está dispoñible en http://corpus.cirp.es/corgaetq e está conformado por un subconxunto de 1197 noticias do

[3] Existe unha versión do *Corpus de Referencia do Galego Actual etiquetado automaticamente* (CORGAetqa), mais polo momento é só interna. Contamos, non obstante, con que estea en breve dispoñible para a súa consulta en liña.

xénero xornalístico e mais por 476 fragmentos textuais, de como máximo 1000 palabras cada un, procedentes da extracción aleatoria de parágrafos de coleccións de relatos.

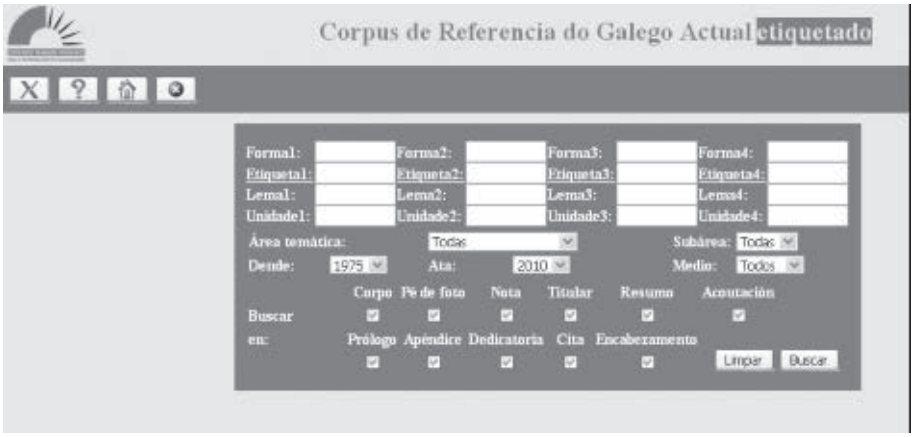
A semellanza da do CORGA, a aplicación de consultas do CORGAetq organízase en varias seccións ás que se accede premendo en cada unha das denominacións:



Contacto, Xestión de usuarios e *Axúdenos a mellorar!* non difiren das do CORGA, mentres que as *Ligazóns de interese* inclúen nesta ocasión corpus etiquetados ou lematizados das linguas peninsulares. Así mesmo, as *Preguntas máis frecuentes* e o *Equipo de traballo* son os específicos para este corpus etiquetado.

Polo que respecta ás *Frecuencias*, nesta sección pódense consultar ou descargar os mil elementos gramaticais máis frecuentes, os mil lemas máis frecuentes ou as frecuencias das etiquetas.

A aplicación de consulta do CORGAetq engade á aplicación de consulta do CORGA a posibilidade de buscar secuencias de ata, como máximo, catro unidades léxicas, etiquetas morfosintácticas e/ou lemas combinados do xeito que se desexe, e pódese ademais facer referencia á unidade gráfica á que pertence cada un dos elementos (elementos Unidade):



Para entender o funcionamento do sistema de consultas é esencial diferenciar entre os termos *unidade*, *forma* e *lema*.

O *lema* é o representante canónico dos elementos que se encadran baixo un mesmo paradigma. Por exemplo, o lema *dar* acolle toda a conxugación verbal do verbo *dar*, de xeito que se se realiza a consulta polo *lema* o sistema devolve as ocorrencias de todos os elementos que se engloban baixo ese paradigma verbal.

A *unidade* identifícase, en xeral, con forma ortográfica, salvo naqueles casos en que a agrupación de máis dunha palabra dá lugar a unidades multipalabra, como son os diversos tipos de locución. Así, ao lado de *chuvia*, *felices*, *fixemos*, *Xan*, *pouquiño*, etc. e dos diversos tipos de amálgamas coma *desta*, *coas*, *dáche* ou *tróuxenllelo*, constitúen tamén unha *unidade*, por exemplo, *Santiago de Compostela*, *San Cibrao das Viñas*, *a carón da* ou *vinte e sete*.

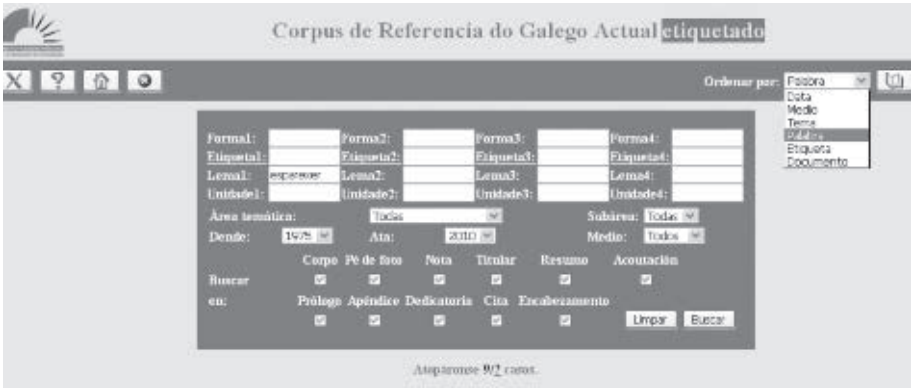
A existencia das amálgamas, ou sexa, de dous ou máis elementos gramaticais baixo unha mesma palabra gráfica onde cada un dos compoñentes posúe unha etiqueta propia e un lema diferente forzan a aparición do elemento *forma*, que identificamos con *unidade léxica* ou *elemento gramatical*; é dicir, a verdadeira unidade de análise do etiquetador, á que lle corresponde sempre unha etiqueta e un lema.

A diferenza percíbese claramente cun exemplo. *Deullo* é unha *unidade* composta pola *forma* 'deu', cuxa etiqueta correspondente 'Vei30s' indica que é a 3ª persoa do singular do pretérito de indicativo do *lema dar*, e mais pola *forma* 'lle' —etiquetada como pronome átono de terceira singular, masculino ou feminino en caso dativo (Rad3as), cuxo *lema* é 'lle'— e, finalmente, pola

forma 'o' —etiquetada como pronome átono de terceira singular masculino en caso acusativo (Raa3ms) cuxo *lema* é 'o'—.

É dicir, a unidade ortográfica *deullo*, non identificable cunha única categoría gramatical nin asociada como un todo a ningunha etiqueta ou lema reais, constitúe en realidade tres elementos *forma* ou *unidades léxicas* con cadanseu par *etiqueta* e *lema* propios. Por isto, cando falamos do tamaño do corpus etiquetado proporciónanse os dous datos: número de palabras ortográficas (617.042) e número de elementos gramaticais (741.833).

Este sistema de consultas permite un amplo abano de buscas, sendo a máis simple a consulta por forma ou lema:



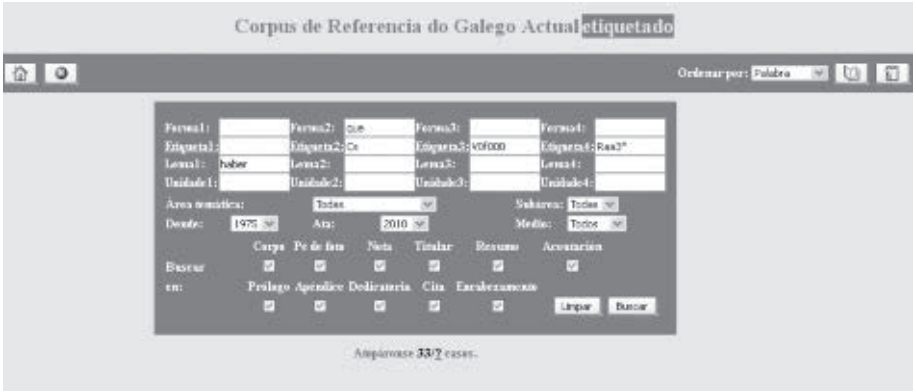
A imaxe amosa a consulta para o lema *esparexer* co que se obteñen todas as formas verbais do paradigma de *esparexer* presentes no corpus. Os resultados pódense ordenar por *data, medio, tema, palabra, etiqueta* ou *documento* e, igual ca no CORGA, dende esta pantalla accédese, a través dos botóns situados á dereita na parte superior da pantalla, ás ocorrencias e ás estatísticas, ou pódese consultar o tamaño do subcorpus sobre o que realizamos a consulta, accedendo deste xeito á nómina de autores e obras, premendo no interrogante que acompaña a casos.

Se observamos a pantalla das ocorrencias, imaxe seguinte, vemos que, de arriba a abaixo, os resultados se visualizan ofrecendo por liña:

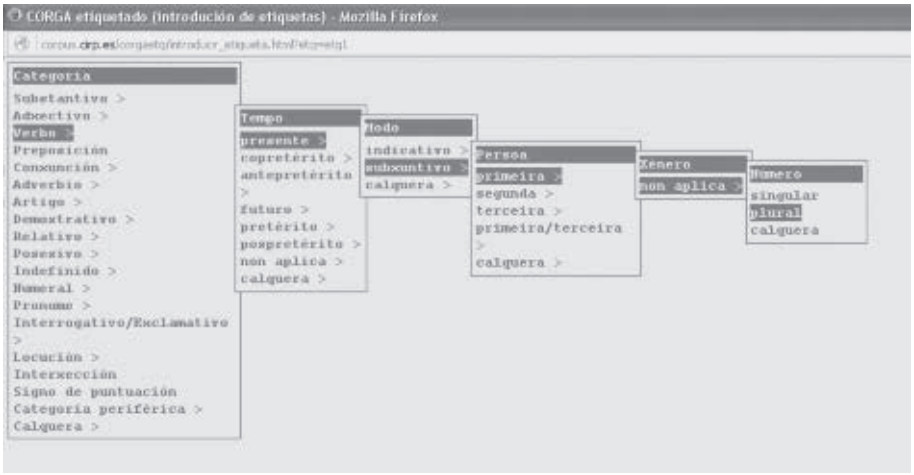
1. Información sobre a localización bibliográfica de cada caso.
2. Número de orde da ocorrencia e mais forma, etiqueta e lema da consulta realizada destacados. Premendo no número accédese ao contexto inmediato da secuencia.

3. Secuencia de análise completa escrita seguindo as normas ortográficas convencionais.
4. Transformación das unidades ortográficas en elementos gramaticais ou formas. Así, a unidade *repartíndose* da liña anterior desagrégase nas formas *repartindo* e mais *se*. En vermello *esparexendo* por corresponder coa busca realizada.
5. Etiquetas correspondentes a cada unha das formas, entre elas, destacada en vermello a de xerundio (V0x000) pertencente ao lema consultado.
6. Finalmente, a liña correspondente aos lemas de cada un dos elementos gramaticais constitutivos da secuencia, destacando en vermello o consultado.

A seguinte imaxe mostra como a etiquetaxe de corpus enriquece considerablemente as posibilidades de buscas ao permitir non só as buscas concretas por forma ou lema, senón tamén a abstracción gramatical a través das etiquetas. A esta riqueza débeselle sumar aínda a posibilidade de utilizar os comodíns que xa vimos para CORGA. A consulta seleccionada é unha das tres que deben realizarse para obter a frecuencia de uso da colocación do pronome átono de terceira acusativo na perífrase de obrigatoriedade con *haber que + infinitivo*. Nesta caso búscase que o pronome apareza ao final.



Unha das vantaxes da aplicación de consulta é que se desenvolveu un menú amigable de consulta por etiqueta que facilita as pescudas cando o usuario non está familiarizado co etiquetario que se emprega. Para acceder a el basta con premer na denominación correspondente á Etiqueta do campo de busca que se desexe —aparece subliñada—; esta acción leva ao menú que se observa na seguinte imaxe, no que o usuario conforma a etiqueta escollendo primeiro a categoría gramatical e, logo, os valores dos atributos correspondentes á clase escolleita, en función das súas necesidades:



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 trátanse como unidades multipalabra ou como unidades consecutivas independentes. Por exemplo, se en *forma* se

busca *dentro*, só van aparecer os casos de *dentro* etiquetados como adverbio, pero non os de *dentro d*e de tipo temporal caracterizados como locución de tipo prepositivo (tipo *Vémonos dentro dun mes*), para cuxa obtención habería que completar en lema *dentro de*, pois a consulta por *dentro de* na *forma* deixaría fóra os casos nos que a preposición *de* está en amálgama.

Para rematar coa descrición do sistema de consultas, cómpre aclarar que por si só *unidade* non é un campo de consulta válido e, polo tanto, 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 das *unidades léxicas* que están amalgamadas. Por exemplo, para comprobar en que casos se emprega o trazo coa denominada segunda forma do artigo basta con especificar:

- Etiqueta1: Dd*
- Lema1: o
- Unidade1: *-l* ou Unidade1: *-*

Se nesta busca en Unidade1 cubrimos só con *l*, consulta que sería lóxica se o elemento *unidade* se correspondese con unidade gramatical, aparecerían 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 unidades léxicas nas que figura un '1', por exemplo *en relación coa*, *por culpa dos* ou *analiza-los*, pois cumpren cos criterios da busca que estamos realizando: 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 *forma*, *etiqueta* e/ou *lema* e recorrer só a *unidade* para restrinxir as buscas, tendo sempre presente que *unidade* non equivale a unidade gramatical.

Por último, cómpre salientar que están dispoñibles para a súa descarga no enderezo http://corpus.cirp.es/xiada/descargas.html os seguintes recursos xerados no proxecto:

- Léxico empregado polo Etiquetador/Lematizador do Galego Actual (Rojo *et al.*, 2015). As entradas, recollidas baixo o formato palabra\tetiqueta\tlema, presentan indicación sobre a súa normatividade, o que fai este léxico indicado tanto para labores de recoñecemento, análise textual e tradución automática como de corrección.

- Corpus de adestramento do Etiquetador/Lematizador do Galego Actual (Rojo *et al.*, 2015). Este corpus inclúe oracións extraídas de xornais, revistas e coleccións de relato curto analizadas seguindo o formato palabra\tetiqueta\tlema. Así mesmo, presentamos unha versión do corpus de adestramento en formato XML na que, amais de incluír a análise dos elementos gramaticais en *forma*, *etiqueta* e *lema*, achegamos tamén a unidade que se analiza, co que se posibilita a recuperación das formas ortográficas simples, multipalabra, amalgamadas ou contraídas e facilítase ao tempo a conexión entre a unidade analizada e as análises correspondentes.

4. Historial de versións

En outubro de 2001 o CORGA ponse, por primeira vez, dispoñible para a consulta a través da rede. Esta primeira versión experimental, constituída por 12,5 millóns de formas, presenta unha codificación mínima e unha cabeceira sinxela que permiten a recuperación da información consonte os parámetros *ano, medio* e *área temática*. En xaneiro de 2003 actualízase esta versión incrementando o número de formas ata os 17,5 millóns.

Ante as limitacións que impón o sistema empregado, vese a necesidade de buscar outro sistema que permita novas funcionalidades e maiores posibilidades na recuperación de información. Decídese entón exportar todo a un estándar XML e así, dende 2004 está dispoñible un novo sistema para o CORGA que utiliza unha codificación XML para os documentos e que é o que evoluciona no tempo (Barcala, 2010). Esa primeira versión XML, constituída por 7 millóns de formas ortográficas, coexiste coa versión do 2003 formada polos 17,5 millóns de formas, posto que non se completara aínda a conversión dos documentos que naquela altura constituían a versión dispoñible na rede.

No 2007 disponse unha nova versión do CORGA que contén case 20 millóns de formas e que xa inclúe todos os documentos da versión antiga e outros moitos novos. Ademais, esta nova versión inclúe novas e significativas posibilidades de busca: consulta da nómina de autores e obras, interconexión do sistema de buscas co sistema de nómina, etc. o que posibilita ter uns valores de referencia moi útiles á hora de extraer conclusións a partir da análise de resultados.

No 2008 publícase a versión 1.3, que contén 23 millóns de formas ortográficas.

No 2009 publícase a versión 1.4, con 25 millóns de formas. Ademais, ponse a disposición pública a versión 1.0 dun sistema máis avanzado de buscas, o *Corpus de Referencia do Galego Actual etiquetado* (CORGAetq), que consta de 250.000 formas ortográficas etiquetadas e lematizadas. Este subcorpus etiquetado, destinado a converterse en corpus de adestramento, está constituído por noticias xornalísticas de temática económica extraídas aleatoriamente do CORGA, cuxa etiquetaxe foi supervisada por unha lingüista.

No ano 2010 publícase a versión 1.5 do CORGA, con 25,8 millóns de formas, e a versión 2.4[4] do sistema de buscas do CORGAetq que inclúe 360.000 formas ortográficas, correspondéndose estas co corpus de adestramento xornalístico.

No ano 2013 publícase a versión 1.6, con 29 millóns de formas, e a versión 2.5 do sistema de buscas do corpus etiquetado que inclúe 499.000 formas ortográficas, as correspondentes ao corpus de adestramento xornalístico e a parte do de ficción.

No ano 2015 publícase a versión 1.7, con 31,9 millóns de formas, e a versión 2.6 do sistema de buscas do corpus etiquetado que inclúe 617.042 formas ortográficas, as pertencentes ao corpus de adestramento xornalístico e de ficción.

En síntese, a listaxe de versións de cada un dos sistemas de buscas, xunto co ano da súa publicación e o número de formas que contén, é a seguinte:

CORGA: Corpus de Referencia do Galego Actual (http://corpus.cirp.es/corga)

versión 1.0, 2001; 12,5 millóns de formas gráficas (textos en txt)

versión 1.1, 2003; 17,5 millóns de formas gráficas (textos en txt)

versión 1.2, 2004; 17,5 millóns de formas en txt (http://corpus.cirp.es/corga) e 7 millóns de formas en xml (http://corpus.cirp.es/corgaxml)

versión 1.3, 2007; 20 millóns de formas gráficas

versión 1.4, 2009, 25 millóns de formas gráficas

versión 1.5, 2010, 25,8 millóns de formas gráficas

versión 1.6, 2013, 29 millóns de formas gráficas

versión 1.7, 2015, 31,9 millóns de formas gráficas

4 O salto na versión da 1.0 á 2.4 prodúcese para igualar co número de versión do etiquetador que se usa, mais non hai outras versións intermedias.

CORGAetq: Corpus de Referencia do Galego Actual etiquetado (http://corpus.cirp.es/corgaetq)

versión 1.0, 2009; 250.000 formas ortográficas / 309.000 elementos gramaticais

versión 2.4, 2010; 360.000 formas ortográficas / 426.000 elementos gramaticais

versión 2.5, 2013; 499.000 formas ortográficas / 594.000 elementos gramaticais

versión 2.6, 2015; 617.000 formas ortográficas / 741.000 elementos gramaticais

5. Liñas de traballo futuras

A incorporación de documentos pertencentes ao novo medio *Internet* (blogs) así como ao novo soporte *Gravación* (transcricións) pon de manifesto a necesidade de adaptar o sistema de consultas para darlles cabida, mais evidencia tamén a mestura na clasificación dos documentos do CORGA entre tipoloxía de documento e tipoloxía temática. Preséntasenos, pois, a oportunidade de reclasificar os documentos discriminando entre estes dous parámetros: tipo de documento e área temática. Facilitaráselle así ao usuario non familiarizado co corpus unha clasificación dos documentos contidos nel coherente e sinxela, manexable. Facilitarase, así mesmo, a consulta por grandes bloques: *ficción* vs. *non ficción*, ou *prensa* vs. *ensaio*, por exemplo, para as que ata agora hai que realizar varias consultas e que só o usuario que coñece ben o sistema é capaz de obter.

Os cambios que terán lugar serán, esquematizados, os seguintes:

En *Medio* inclúense os valores *internet* (clasificará os blogs) e *audiovisual* (clasificará os guións e as transcricións).

En *Soporte* engádese *gravación* para acoller as transcricións. Na recuperación de información poderase escoitar o audio correspondente ao contexto da busca realizada.

En *Área temática* desaparecen os valores correspondentes á *Ficción*, que se transformarán en valores de *tipoloxía textual*, igual que os que se escollan para clasificar as diferentes transcricións (entrevista, informativo, discurso, conversa etc.).

O menú de consulta, *grosso modo*, irase abrindo a medida que elixamos posibilidades de busca. Así, seguirá os seguintes pasos:

1. Tipo de documento: *todo, escrito, oral*

   Se se opta por *todo* a consulta será a todo o corpus.

   Se se preme en *oral* accederase directamente á tipoloxía na que se clasifiquen as transcricións (entrevista, informativo, conversa, discurso etc.), ademais de a un valor *todo.*

   Se se opta por *escrito*, haberá unha segunda clasificación en *ficción* e *non ficción.*

2. *escrito*: *todo, ficción, non ficción*

   todo: a consulta será a todos os documentos escritos do corpus. Non poderá especificarse área temática.

   ficción: poderase escoller entre *todo* ou cada un dos tipos de documento incluídos en ficción: *novela, relato, teatro, guión.*

   non ficción: poderase escoller entre *todo, prensa, ensaio, blog*. A escolla de *prensa* poderá, á súa vez, desagregarse en *xornal* e *revista.*

É dicir, as áreas temáticas e subáreas habilitaranse unicamente para o valor *non ficción* ou calquera dos valores contidos na *non ficción.*

Outro dos obxectivos destacables a curto prazo, ademais de ampliar as funcionalidades de busca para facilitar a detección de colocacións e mellorar as posibilidades de visualización con gráficos, é a unificación dos dous sistemas de consulta dispoñibles actualmente en liña (CORGA e CORGAetq), xunto co sistema do CORGA etiquetado automaticamente que de momento é só de uso interno, nun único sistema que dea cabida ás diferentes aproximacións de busca: consulta por forma ortográfica ou consulta por forma, etiqueta e/ou lema no subcorpus desambiguado manualmente ou no corpus etiquetado automaticamente.

A estas melloras engadiráselles, por último, outra non menos importante cualitativamente para potenciar as funcionalidades de busca do corpus etiquetado. Trátase da implementación no sistema de consultas dun campo máis, o de *hiperlema*, ao lado do de *forma, etiqueta* ou *lema*. Co seu emprego, optativo en función dos intereses concretos de quen realiza a consulta, o

usuario do corpus etiquetado poderá abstraerse da variación ortográfica existente nos textos, entendida esta como a modificación gráfica que sofre unha forma respecto da forma canónica correspondente coa que mantén unha relación clara de identidade. Poñamos por caso, se alguén quixese estudar o esquema argumental do verbo *comezar*, agradecería recuperar cunha soa consulta todos os exemplos dos lemas *comezar, comenzar, escomezar, escomenzar, encomezar* e *encomenzar*; pois ben, isto será posible na próxima versión do CORGA etiquetado, manual ou automaticamente, coa consulta sobre o hiperlema *comezar*.

Bibliografía

Mair, C. (2006): «Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora», en A. Renouf e A. Kehoe, *The changing face of corpus linguistics*, Amsterdam: Rodopi, pp. 355-376.

Barcala Rodríguez, F. M., M. A. Molinero e E. Domínguez (2005): «Construcción de sistemas de recuperación de información sobre córpora textuales estructurados de grandes dimensiones», *Procesamiento del Lenguaje Natural* 34, pp. 41-48.

Barcala Rodríguez, F. M (2010): *Corpus lingüísticos estruturados de grandes dimensións: Metodoloxía e sistemas de recuperación de información*. Tese de doutoramento, Universidade da Coruña, <http://ruc.udc.es/dspace/bitstream/2183/7171/1/tese_mario_barcala.pdf>

Domínguez Noya, E. Mª. (2008): «O Corpus de Referencia do Galego Actual (CORGA): presente e futuro», en E. González Seoane, A. Santamarina e X. Varela Barreiro, *A lexicografía galega moderna. Recursos e perspectivas*, Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega, pp. 139-151.

Domínguez Noya, E. Mª., F. M. Barcala Rodríguez e M. Á. Molinero Álvarez (2009): «Avaliación dun etiquetador automático estatístico para o galego actual: Xiada», *Cadernos de Lingua* 30/31, pp. 151-193.

Domínguez Noya, E. Mª. (2013): *Etiquetaxe e desambiguación automáticas en galego: o sistema XIADA*. Tese de doutoramento, Universidade de Santiago de Compostela: Repositorio Institucional da USC <http://hdl.handle.net/10347/9587>

Rojo, G., M. López Martínez, E. Domínguez Noya e F. M. Barcala (2015): *Listado de frecuencias do Corpus de Referencia do Galego Actual (CORGA),*

versión 1.7, Centro Ramón Piñeiro para a Investigación en Humanidades, <http://corpus.cirp.es/corga/frecuencias.tar.gz>.

Rojo, G., M. López Martínez, E. Domínguez Noya e F. M. Barcala (2015): *Corpus de adestramento do Etiquetador/Lematizador do Galego Actual (XIADA), versión 2.6*, Centro Ramón Piñeiro para a Investigación en Humanidades, <http://corpus.cirp.es/xiada/corpus_xiada_2_6.tar.gz>

Rojo, G., M. López Martínez, E. Domínguez Noya e F. M. Barcala (2015): *Léxico do Etiquetador/Lematizador do Galego Actual (XIADA), versión 2.6*, Centro Ramón Piñeiro para a Investigación en Humanidades, <http://corpus.cirp.es/xiada/lexico_xiada_2_6.tar.gz>