CORGA
(Corpus de Referencia del Gallego Actual)

María Sol López Martínez

Centro Ramón Piñeiro (CRPIH) / Universidade de Santiago de Compostela

El *Centro Ramón Piñeiro para a Investigación en Humanidades* (CRPIH) tiene, entre sus objetivos, el desarrollo de recursos informáticos que den salida a proyectos de investigación nacidos en el propio Centro y que ayuden a la incorporación de la lengua gallega al nuevo mundo de las tecnologías de la información.

El proyecto CORGA[1] nace con la idea de poner a disposición de la comunidad científica un nuevo recurso que pueda ser accesible a través de Internet. La finalidad del CORGA es, por tanto, facilitar la obtención de datos para el estudio de aspectos morfológicos, sintácticos y léxicos de la lengua gallega para un período determinado, como veremos.

El CORGA es una colección de textos lingüísticos y reales, almacenados en formato electrónico, seleccionados según unos determinados criterios y con la suficiente amplitud para ser considerados representativos del uso lingüístico del gallego actual. Quiere ser un Corpus de Referencia del Gallego Actual y, en consecuencia, además de tener un cierto grado de amplitud, deben estar representados los diferentes tipos de textos de la lengua gallega actual. Para lograr esa representatividad se han definido los criterios que, a nuestro entender, se deben de tener en cuenta para la selección de los textos, y se ha diseñado la configuración que, desde nuestro punto de vista y teniendo en cuenta las características de la lengua gallega, debe conformar el corpus.

En la elaboración de un corpus diferenciamos cuatro partes: el diseño y la configuración de los documentos que deben formar de él, la introducción de los textos y su codificación, la anotación y etiquetación y por último la utilización. En

[1] Dirigen el proyecto Guillermo Rojo (Director Técnico del Área de Lingüística del CRPIH), María Sol López Martínez y Francisco García Gondar. Hasta este momento han participado en su construcción los siguientes becarios: M. Teresa Araújo García (1994-2001), Cristina Blanco González (2001-), Inés Diz Gamallo (1994-2001), Eva Domínguez Noya (1995-), Beatriz Fernández Paredes (2000-), Susana Ferreiro García (1998-2001), Susana García Rodríguez (1997-2000), Deborah González Martínez (2001), Ana Ledo Villaverde (2000-), Mónica Martínez Baleirón (1998-2000), M. Teresa Monteagudo Cabaleiro (1994-1998), Xesús Mosquera Carregal (2000- 2002), Luísa Pita Rubido (2001-.), Antón Porto Sánchez (1998-1999), Sonia Varela Pombo (1998-2001) y Pilar Vázquez Grandas (1994-1998). Los responsables de la parte informática son Francisco Mario Barcala (2000-), Jesús Rodríguez Castro (1995-1998), Fernando Magán Muñoz y José Carlos Sánchez Rivas. El Sistema de Búsquedas lo desarrolló el *Instituto de Investigacións Tecnolóxicas* de la USC (IIT).

esta parte me referiré a las dos primeras partes, es decir, codificación y diseño e introducción y codificación de los textos.

Como señalamos anteriormente, un corpus de referencia debe contener una muestra equilibrada de los distintos tipos de textos, ello implica la definición de aquellas características que, entendemos, deben tener los textos seleccionados. Así pues, desde nuestro punto de vista y teniendo en cuenta las características de la lengua gallega, en el diseño del CORGA se tuvieron en cuenta los siguientes parámetros: Tamaño, Medio, Áreas temáticas e Representatividad de los períodos.

Cronológicamente el CORGA se concibió como una muestra representativa de la lengua gallega de los últimos 30 años. En esta fase, el CORGA incorpora textos publicados entre 1975 y 2004 y al final su tamaño será de 25 millones de formas. El CORGA está constituido en su mayor parte por textos escritos y por una pequeña parte de textos orales (5% del total del corpus). De los primeros se seleccionaron textos de ficción (narrativa y teatro); y de no ficción (ensayos, manuales, revistas y periódicos).

Según el medio de publicación diferenciamos entre libros, revistas y periódicos. La aportación de cada uno de ellos al conjunto será el siguiente:

Libros	60%
Revistas	20%
Periódicos[2]	20%

Los documentos se clasifican por su contenido. Aquí diferenciamos 6 áreas temáticas, además del oral, con la siguiente representatividad para cada una de ellas:

Economía y política	15%
Cultura y artes	15%
Ciencias sociales	15%
Ciencia y Tecnología	10%
Otros	15%
Ficción	30%

La representatividad a que hemos hecho referencia anteriormente debe reflejarse también en la distribución de los textos por años o períodos. Dado que la finalidad del corpus es presentar una muestra de la lengua del momento, entendemos que los porcentajes de cada uno de los años o períodos no debe ser igual. Así pues, para establecer la representación cronológica los agrupamos por lustros, procurando dar siempre una mayor presencia a los períodos más recientes. La propuesta de distribución se puede ver a continuación:

1975-79	5%
1980-84	10%
1985-89	15%

[2] En el CORGA los textos procedentes de periódicos se incorporan a partir de 1995, año en el que se empieza a publicar en gallego *O Correo Galego.*

1990-94	20%
1995-99	25%
2000-04	25%

En definitiva de acuerdo con los parámetros que hemos comentado, los textos seleccionados se clasifican por la *Fecha de publicación*, *Medio* y *Área temática*. Estos tres criterios son los que se podrán utilizan para la recuperación de información. De este modo, el sistema devolverá datos y ejemplos sobre una forma en todo el corpus, o se podrá restringir a uno o más de los parámetros anteriores.

Una vez hecho el diseño, procedemos a la selección de los textos para posteriormente incorporarlos al corpus. Los documentos que incorporamos al CORGA proceden de ediciones impresas, versiones electrónicas independientes y grabaciones de textos orales. No obstante, la mayor parte de ellos proviene de ediciones impresas. Únicamente en los años más recientes se incorporaron algunos documentos procedentes de internet. Es el caso, por ejemplo, de la versión electrónica del periódico *O Correo Galego* o de otro tipo de documentos tomados también de la red. Por último, para los textos orales se han recogido grabaciones de programas de los medios audiovisuales.

En la configuración del corpus tuvimos que resolver algunos problemas derivados de la situación de la lengua. Estos problemas se centran fundamentalmente en tres aspectos:

a) Normativos. Los responsables del proyecto CORGA entendemos que, desde la aprobación en 1982 de las *Normas ortográficas e morfológicas do Idioma Galego* elaboradas por el Instituto da Lingua Galega (ILG) y la Real Academia Galega (RAG), los textos seleccionados para formar parte del corpus deben cumplir la normativa. Aunque mayoritariamente ocurre así, no podemos olvidar que el CORGA incorpora documentos publicados a partir del año 1975, por tanto, anteriores a la normativa. En los textos de estos períodos es habitual la presencia de variantes gráficas y también morfológicas. Pero, incluso, en aquellos posteriores a la aprobación de la normativa los autores utilizan variantes en unos casos de carácter gráfico, en otros morfológico o incluso léxico. Con todo, a pesar de algunas diferencias (en el respeto a la normativa) entre los textos, en el CORGA se puede constatar una mejora notable en el proceso de normativización de la lengua.

b) Dificultades para cubrir algunas áreas temáticas. Se puede afirmar que hoy en día podemos encontrar publicaciones en gallego sobre prácticamente todos los temas. Sin embargo, en algunas áreas se hace un poco más difícil conseguir el porcentaje de formas previsto en el diseño general. Es el caso, por ejemplo, de las áreas de Ciencias y Tecnología.

c) Prensa. La falta de prensa diaria hasta 1995 hace que exista un cierto desequilibrio en esta clase de textos en los primeros períodos. Aunque otro tipo de publicaciones periódicas ya existían con anterioridad, hasta 1995 no fue posible incorporar documentos procedentes de periódicos. Y además hasta este año la parte de prensa estaba cubierta únicamente por *O Correo Galego*; a partir de aquí se incorporarán también documentos procedentes del periódico semanal *A nosa Terra*.

El CORGA actual

Aunque al CORGA se siguen incorporando documentos de acuerdo con el diseño propuesto, en este momento ya está disponible para su consulta en la web (http://corpus.cirp.es/corga). Esta versión está constituida únicamente por textos escritos y tiene un tamaño de aproximadamente 17,5 millones de formas gráficas; de estas 329.803 son distintas. Además en la información sobre la frecuencia de las formas se puede constatar que, entre las 50 más frecuentes del corpus, la mayor parte de ellas son palabras gramaticales, y, sólo, a partir del puesto 30 aparecen algunas formas del verbo 'ser'. En esta versión ya se pueden obtener datos de todos los períodos, no obstante se hace necesario completar el último lustro y al mismo tiempo proceder a la implementación de nuevos materiales de otros períodos para cumplir con los criterios de configuración diseñados previamente.

El proceso de implementación de nuevos documentos al CORGA se hace siguiendo los criterios y la clasificación con los que se diseñó la configuración del corpus. Las nuevas incorporaciones las hacemos por bloques de 2,5 millones de formas. De esta manera, después de cada implementación se puede comprobar la distribución de los textos teniendo en cuenta los criterios propuestos. Con este método de trabajo, si se detectan desvíos significativos en alguno de los parámetros, estos podrán corregirse en la próxima ampliación.

Los documentos que forman parte del CORGA proceden siempre de textos publicados. Si existe más de una edición, siempre se toma como referencia la primera que será la que sirva de referencia para datar la lengua utilizada. Generalmente, los textos seleccionados se incorporan íntegramente al corpus; sólo en algunos casos, y para cubrir el contenido de algunas áreas temáticas y, al mismo tiempo, evitar desequilibrios entre ellas, se han seleccionado partes de publicaciones periódicas.

Cada documento está formado por una cabecera y el texto propiamente dicho. En la cabecera se indican los datos bibliográficos (autor, título, año de publicación, lugar, depósito legal, ISBN, etc., el tipo de documento (libro, revista, periódico) y el área o áreas temáticas en las que se incluye. También forman parte de la cabecera otros datos como el número de palabras, el número de bytes, el responsable de la codificación, etc. En el cuerpo del documento la codificación es mínima. Se marca(n) aquella(s) parte(s) del documento que aparece escrita(s) en otro idioma; esto quiere decir que en la recuperación de información no podrán aparecer formas de otras lenguas.

En el diseño y configuración de corpus hemos propuesto unos porcentajes en cada uno de los parámetros que utilizamos para la recuperación de información. Naturalmente, esa propuesta deberá cumplirse al término de esta fase. En el CORGA que presentamos aquí la distribución en algunos de los parámetros presenta un cierto grado de desviación, no obstante, nos parece importante presentarla para que se tengan en cuenta al trabajar con los datos de esta versión:

a) Medio

Libros	66,80%
Periódicos	26,41%
Revistas	6,78%

b) Períodos

1975-79	4,75%
1980-84	8,11%
1985-89	9,27%
1990-94	15,39%
1995-99	44,40%
2000-04	17,98%

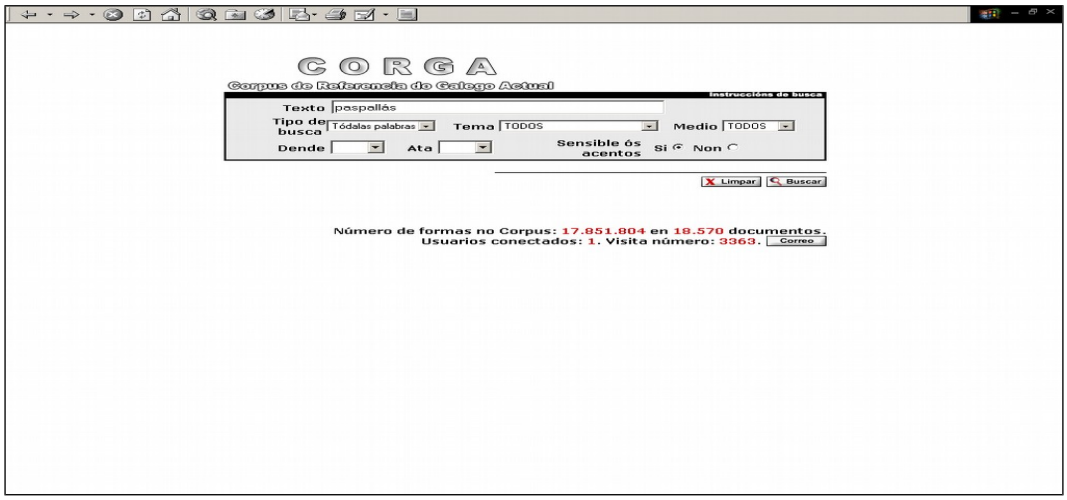
c) Área Temática:

Economía y Política	22,5%
Cultura y Artes	6,80%
Ciencias Sociales	13,32%
Ciencias y Tecnología	5,67%
Otros	11,96%
Ficción	39,79%

Queremos hacer notar que del último período (2000-04) hasta este momento sólo hemos incorporado textos publicados en el 2000 y 2001. Las siguientes ampliaciones hasta los 25 millones previstos deben permitir la redistribución de los materiales, de acuerdo con la configuración prevista, para ajustarse lo más posible al diseño general del corpus.

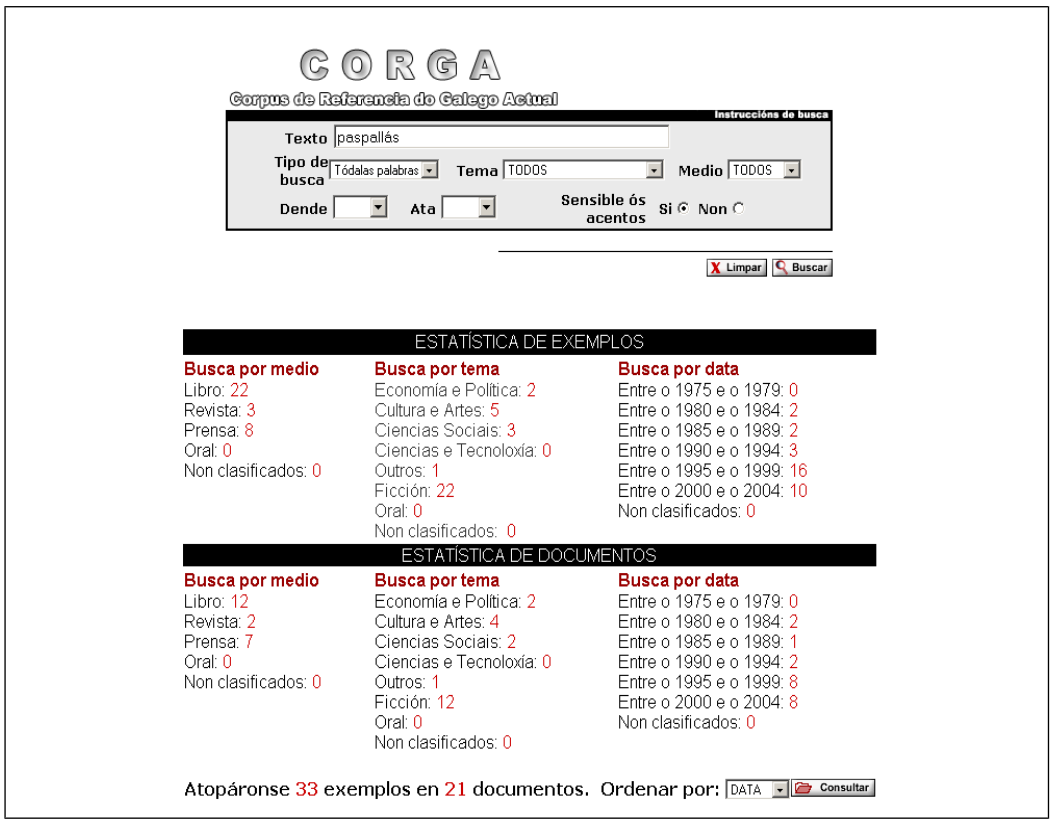
La explotación del CORGA que está en la red a disposición de los investigadores presenta algunas limitaciones en la recuperación de información. Estas se deben al poco nivel de codificación y, sobre todo, a la falta de anotación. No obstante, nos parece un recurso con información muy importante para poder afrontar estudios de carácter gramatical sobre la lengua gallega actual. El sistema actual de recuperación de información permite hacer búsquedas de formas ortográficas, expresiones regulares, mediante la utilización de los comodines asterisco (*) e interrogación (?). También se pueden utilizar los operadores booleanos (AND, OR, AND OR). Pero naturalmente, no es posible en este CORGA hacer búsquedas por lemas o por categoría gramatical.

A continuación veremos un ejemplo de búsqueda y la información que se muestra en las diferentes pantallas. Así pues, el usuario, una vez dado de alta, accede a la pantalla de búsquedas.



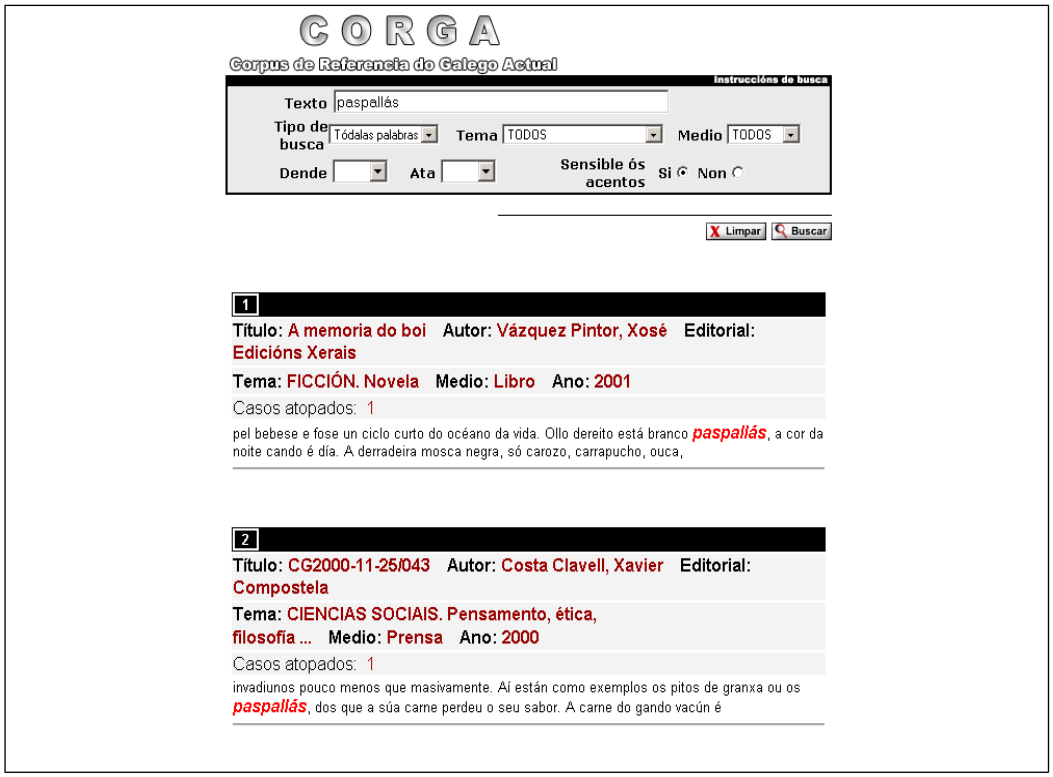
En ella se muestran varias posibilidades de búsqueda (Todas las palabras, cualquier palabra, frase exacta, o con operadores booleanos). El tipo de búsqueda se puede referir a todos los períodos del corpus o marcar unos años determinados. También se pueden combinar con la clasificación temática y/o por medio.

Una vez hecha la solicitud, el sistema nos devuelve dos pantallas de presentación de datos. En la primera se puede ver el número de ejemplos y de documentos en los que aparecen la forma o formas buscadas.



En esta imagen se puede ver que la información aparece clasificada además por los parámetros *Medio, Fecha y Tema*. Y también se nos dice que en todo el corpus hay 33 ejemplos de esta forma y estos se encuentran distribuidos en 21 documentos.

En las pantallas siguientes, el usuario puede ver los documentos en los que aparece la búsqueda solicitada. En cada una el sistema nos presenta un total de 10 documentos que pueden aparecer ordenados por fecha, medio o área temática. Además en cada documento se indican los datos de la cabecera que lo identifica bibliográficamente y los valores de clasificación que se le asignaron, el número de casos encontrados y el contexto en el que aparece. Como se puede observar en la imagen que viene a continuación, en cada ejemplo la(s) forma(s) solicitada(s) aparece(n) destacadas en color rojo.



El acceso para la utilización del CORGA es gratuito, mediante la obtención previa de una licencia de uso. Para ello el solicitante debe seleccionar en la primera pantalla la opción 'Rexistro de usuarios', cubrir el formulario y enviarlo. El sistema le devolverá un nombre de usuario y una contraseña para poder acceder al corpus.

Desarrollos previstos en el CORGA

Como hemos visto anteriormente, la recuperación de información en la versión actual del corpus es, en algunos aspectos, limitada debido a un bajo nivel de estructuración de los documentos y, naturalmente, a la ausencia de anotación. La mejora de estos aspectos la estamos afrontando en dos direcciones distintas, pero al mismo tiempo complementarias.

Por un lado, hemos comenzado la revisión de los documentos que forman parte del CORGA para proceder a una codificación más exhaustiva y posteriormente archivarlos en formato .xml. A diferencia del formato anterior en este un periódico,

una revista, una colección de relatos o de ensayos se consideran un único documento que estará constituido por una cabecera y un contenido. El contenido de los documentos de este tipo está constituido a su vez por un número determinado de noticias, relatos o de ensayos. Para cada una de las partes (noticia, relato o ensayo) se diferencia la cabecera y el cuerpo, es decir, el texto. Y además en todos los documentos se marcan los parágrafos y las oraciones. Con esta mejora en la codificación tratamos de incrementar las posibilidades del sistema de consulta.

Paralelamente a los cambios en la codificación, se está trabajando en la construcción de un analizador morfológico para, en una fase posterior, poder llevar a cabo la anotación y desambiguación automáticas de todo el CORGA. El analizador está prácticamente acabado, consta de 693 etiquetas, de un lexicón formado por 31.200 lemas y, de hecho, ya se utilizó para anotar aproximadamente unas 100 mil formas que, posteriormente, se desambiguaron manualmente. Para conseguir un buen rendimiento de este recurso informático, tenemos que resolver algunos problemas que plantean la etiquetación de una parte de las formas que hay en el corpus. Anteriormente, hemos señalado que los textos del gallego, sobre todo los anteriores a 1982, presentan una importante variación de carácter gráfico, morfológico e incluso léxico. La eficacia del analizador requiere previamente buscar soluciones para lograr que reconozca y analice automáticamente no sólo las formas normativas, sino también aquellas que presentan algún tipo de variación.