O CORPUS DE REFERENCIA DO GALEGO ACTUAL (*CORGA*): PRESENTE E FUTURO

Eva Mª Domínguez Noya

Centro Ramón Piñeiro para a Investigación en Humanidades

As posibilidades de recuperación de información nun corpus dado increméntanse notablemente en consonancia co seu grao de codificación. Canto máis estruturado e codificado estea este máis información extraemos das buscas. Pénsese que se os textos están marcados exhaustivamente sempre podemos decidir, como usuarios, se nos interesa toda a información resultante da busca executada ou se pola contra nos limitamos a estudar as diferentes ocorrencias da busca obxecto de estudo. Se o texto non presenta marcas de ningún tipo ou estas son mínimas teremos que cinguirnos a buscas moi simples e a información que desexamos ou non a podemos obter ou precisaremos dar moitas voltas para conseguila. Por exemplo, se os textos que integran un corpus non teñen asignada unha área temática que os caracterice será tarefa ardua delimitar os ámbitos nos que unha palabra ou expresión se emprega.

Como saben todos os que nalgún momento consultaron o *Corpus de Referencia do Galego Actual* (*CORGA*), este é un corpus documental que en novembro de 2006 consta duns 13 millóns de formas, integrado por distintos tipos de textos –xornais, semanarios, revistas, ensaios e textos de ficción (novela, relato curto e teatro)–, que abrangue temporalmente dende o ano 1975 ata a actualidade e que están codificados no estándar XML (*eXtensible Markup Language*). E aquí, cando falamos de codificación, referímonos principalmente á información bibliográfica e á estruturación do documento. Exemplificando cun texto xornalístico, o XML permítenos considerar un xornal como un único documento que está organizado en noticias, 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 punto, punto e coma, dous pun-

tos, etc.). Esta disposición detallada habilita a posibilidade de realizar consultas sobre a totalidade do documento (noticia) ou sobre unha unidade estrutural concreta (titular, resumo, pé de foto ou corpo). A codificación inclúe ademais a marcación de fragmentos que aparecen nunha lingua distinta do galego –así evitamos indexar eses fragmentos e impedimos que engorden os datos totais do corpus no referente a número de palabras e tamén que se poidan realizar buscas sobre eles–, os poemas, as entrevistas, as táboas, as fórmulas, etc.

Sobre este conxunto de formas gráficas é posible facer buscas de palabras ou expresións en xeral, por tipos de texto, épocas, áreas temáticas ou calquera combinación dos parámetros anteriores. Proximamente contamos con poñer na rede unha nova versión do *CORGA* que, ademais de chegar aos case 20.000.000 de formas, engade un novo sistema de consulta da nómina de autores e obras que permite buscar que obras ou autores están no corpus, saber que número de palabras totais e documentos corresponde á busca realizada ou que cantidade de palabras contén o *CORGA* nunha certa área temática, período de tempo, etc.

Por exemplo, se buscamos unha forma como "abreconcertos" nos xornais que teñan como área temática principal economía e política, nos resultados da consulta, cando se nos ofrece o número de oracións e documentos atopados de todo o corpus, podemos visualizar o número de palabras e documentos que cumpren as condicións esixidas na consulta inicial.

Somos conscientes, sen embargo, de que para facer buscas máis avanzadas é imprescindible que os textos do *CORGA* estean lematizados e etiquetados, e levamos xa algúns anos traballando nesta dirección, e aquí, cando falamos de etiquetación, debemos aclarar que entendemos esta como o proceso mediante o cal a cada unidade léxica se lle asigna un lema e unha etiqueta. En concreto, no Centro Ramón Piñeiro para a Investigación en Humanidades desenvólvese paralelamente á construción do *CORGA* un proxecto denominado *Etiquetador e lematizador do galego actual* destinado precisamente a lematizar e etiquetar os textos de *CORGA*, ao que internamente coñecemos como XIADA, e sobre o que imos centrar este traballo tal como nos solicitaron os organizadores deste volume.

Á hora de abordar a construción dun analizador e etiquetador cómpre atender as seguintes frontes:

(1) Determinación dun sistema de etiquetas.
(2) Deseño da estrutura do lexicón e construción deste.
(3) O preprocesador.
(4) O etiquetador.

1. DETERMINACIÓN DUN SISTEMA DE ETIQUETAS

Se queremos asignarlle ás formas de *CORGA* unha etiqueta, primeiro deberemos decidir que etiquetas imos empregar, que clases morfolóxicas imos distinguir e que atributos as van caracterizar, e para iso é imprescindible delimitar un etiquetario ou *tagset*.

O etiquetario para o galego foi desenvolvido no marco do proxecto *Construción, etiquetación e lematización do Corpus de Referencia do Galego Actual* como unha ferramenta para a análise computacional e a anotación automática dos textos do *CORGA*. Para a elaboración desta etiquetaxe seguíronse as recomendacións de EAGLES (Leech / Wilson 1996) e os textos básicos da lingüística galega. Así, o sistema de etiquetación de XIADA presenta unha estrutura xerárquica na que no primeiro nivel se identifica a categoría morfolóxica e nun segundo os atributos gramaticais pertinentes que constitúen o conxunto básico de atributos para cada categoría. Esta división en dous niveis percíbese claramente no formato: a letra que identifica a clase de palabra á que pertence o elemento vai en maiúscula e o valor dos atributos represéntase con minúsculas.

Na construción do *tagset* primamos a descrición morfolóxica e reducimos a información sintáctica á caracterización de determinadas categorías cuxos elementos integrantes poden funcionar como determinantes ou non determinantes. Delimitamos, en primeiro lugar, as clases de palabras existentes e a continuación establecemos os atributos gramaticais pertinentes que permitisen o recoñecemento morfolóxico de calquera palabra:

						n-Nominativo **a**-Acusativo				
		m- Masculino				**d**-Dativo	**p**-Presente	**i**-Indicativo		
		f- Feminino	**s**- Singular		**1**-Primeira	**i**-Impersoal	**i**-Copretérito	**s**-Subxuntivo		
		a- Masculino	**p**- Plural		**2**-Segunda	**p**-Preposicional	**l**-Antepretérito	**m**-Imperativo		
		e- Feminino	**a**- Singular	**c**- Comparativo	**3**-Terceira	**r**-Nom./Prep.	**f**-Futuro	**f**-Infinitivo	**s**-Singular	**n**-Non Determinante
		n-Neutro	**e**- Plural	**s**- Superlativo	**a**-Primeira e Terceira	**l**-Acus./Dat.	**e**-Pretérito	**x**-Xerundio	**p**-Plural	**d**-Determinante
		0-Non aplicable	**0**- Non aplicable	**0**-Non aplicable	**0**-Non aplicable	**b**-Acus./Dat./ Imp.	**c**-Postpretérito	**p**-Participio	**a**-Singular e Plural	**a**-Determinante e non Determ.
Categoría	**Tipo**	**Xénero**	**Número**	**Grao**	**Persoa**	**Caso**	**Tempo verbal**	**Modo**	**Posuidor**	**Modo**
Substantivo (**S**) 1	**c**-Común **p**-Propio 2	3	4							
Adxectivo (**A**) 1		3	4	2						
Verbo (**V**) 1		6	5		4		2	3		
Preposición (**P**) 1										
Conxunción (**C**) 1	**c**-Coordinante **s**-Subordinante 2									
Adverbio (**W**) 1	**n**-Nuclear **m**-Modificador **a**-Nucleo e Modificador **r**-Relativo **g**-Int.-Exctivo 2									
Artigo (**D**) 1	**d**-Determinado **i**-Indeterm. 2	3	4							

Demostrativo (**E**) 1		3	4							2
Relativo (**T**) 1		3	4							2
Posesivo (**M**) 1		5	6		3				4	2
Indefinido (**I**) 1		3	4							2
Numeral (**N**) 1	**c** - Cardinal **o** - Ordinal 2	4	5							3
Pronome (**R**) 1	**t**- Tónico **a** - Átono 2	5	6		4	3				
Exclam./ Interr (**G**) 1		3	4							2
Interxección (**Y**) 1										
Sinal de puntuación (**Q**) 1										
Categoría Periférica (**Z**) 1	**f**- Fórmula **a**- Abreviatura **g**- Sigla **s**- Símbolo **o**- Outros tipos 2	3	4							

Como se pode observar no *tagset* de XIADA que acabamos de reproducir (dispoñible en http://corpus.cirp.es/xiada), na primeira columna recóllense as categorías gramaticais e as letras coas cales as denominamos:

- Adverbio W
- Adxectivo A
- Artigo D
- Conxunción C
- Demostrativo E
- Exclamativo-Interrogativo G
- Indefinido I
- Interxección Y
- Locución L
- Numeral N
- Posesivo M
- Preposición P
- Pronome R
- Relativo T
- Sinal de Puntuación Q
- Substantivo S
- Verbo V
- Categoría Periférica Z

Nas seguintes columnas recóllense os atributos e os valores destes e, por último, nas filas incorpórase un número que indica a posición que o atributo, se é pertinente, ocupa dentro da etiqueta final.

A distribución dos atributos dentro de cada categoría é simétrica, de tal xeito que todos os elementos pertencentes a unha categoría dada posúen o mesmo número de atributos e os seus valores ocupan a mesma posición dentro da cadea resultante.

Hai, sen embargo, etiquetas nas que en relación co grao de especificación ao que queiramos chegar, ou ben porque a atribución dun valor inequívoco sexa imposible, podemos neutralizar o valor dalgún ou de varios atributos. Por exemplo, nun texto de prensa, atopámonos coa expresión *perigos e incertezas existentes*, onde non sabemos se o adxectivo concorda co substantivo masculino *perigos*, co feminino *incertezas* ou cos dous. Nestes casos é necesario manter a ambigüidade, polo que a palabra *existentes* se etiqueta como "Adxectivo grao non aplica masculino ou feminino plural" (A0ap). En XIADA marcamos esta ambigüidade co

valor "a", que representará o xénero, número, persoa, posuidor ou valor dependendo da clase de palabra e da posición que ocupe na cadea de etiquetaxe.

Por outra banda, non todos os elementos pertencentes a unha mesma categoría morfolóxica responden positivamente aos atributos que a caracterizan. Así, na categoría verbal, o atributo xénero só é un trazo pertinente para a caracterización do participio. Para o resto de elementos do paradigma verbal a aplicación do atributo xénero non é pertinente. Por este motivo creamos un valor "0" co significado de "Non aplica".

Consideráronse como atributos primordiais os seguintes: tipo, subtipo, xénero, número, grao, persoa, caso, tempo verbal, modo, posuidor e valor. Combinando as clases gramaticais delimitadas e os atributos que as caracterizan, o noso sistema de etiquetación presenta un total de 409 etiquetas posibles.

Poderiamos explicar detalladamente a escolla dos valores de cada un destes atributos e como afectan á clase implicada pero o espazo de que dispoñemos non é moito e cremos que outras fases do proxecto resultan tan interesantes ou máis.

2. DESEÑO DA ESTRUTURA DO LEXICÓN E CONSTRUCIÓN DESTE

Normalmente os etiquetadores necesitan un lexicón onde almacenar a información de etiqueta e lema que lle corresponde a cada palabra. Ademais, aínda que os etiquetadores intentan adiviñar a etiqueta dunha forma que non estea presente no lexicón, canto maior sexa este, máis posibilidades de acerto hai.

Á hora de afrontar a elaboración dun recurso léxico e gramatical calquera destinado a un uso computacional é fundamental que o deseño deste cubra todas as necesidades actuais e, ao tempo, contemple as futuras.

Por unha banda, dado que a introdución e xestión dos datos debe ser cómoda, consideramos que calquera entrada do lexicón debe ser caracterizada morfoloxicamente desagregándoa en *lema*[1], *raíz*[2], *subetiqueta*[3] e

1. Xeralmente correspóndese coa entrada dun dicionario. Por exemplo "neno".
2. Segmento inicial do lema sen desinencia se esta é recorrente. Por exemplo "nen".
3. Definición e caracterización morfolóxica da forma obxecto de análise sempre e cando estea completa e non a enviemos a ningún grupo derivacional.

grupo de derivación[4]. Esta estrutura esixiu un estudo formal detallado dos integrantes de cada unha das clases gramaticais variables para agrupalos en modelos e deste xeito construír a gramática formal que nos permitise analizar e tamén xerar as formas flexionadas e conxugadas do galego moderno. Evitamos así, por exemplo, introducir manualmente as 65 formas pertencentes a calquera verbo e incluímos só as raíces, asignándolles a subetiqueta *Verbo* e enviándoos ao grupo de derivación que corresponda.

Por outra banda, o noso lexicón non só é unha base de datos na que constan os lemas coas respectivas indicacións morfolóxicas de clase gramatical, xénero, número, tempo, modo, etc., senón que tamén nos proporciona máis información sobre cada unha das entradas, información non só morfolóxica senón tamén lingüística.

Así, en primeiro lugar, o recoñecemento e caracterización das formas existentes nos textos reais contemporáneos galegos esixe que o lexicón estea integrado por lemas normativos e non normativos. Se a construción do lexicón se realizase con suxeición á actual normativa ortográfica e morfolóxica da lingua galega, a etiquetación e análise dos textos dun corpus representativo do galego actual como o *CORGA* non sería de utilidade para os investigadores posto que quedarían numerosas formas sen recoñecer automaticamente. Pareceunos interesante, en consecuencia, rexistrar na estrutura do lexicón o parámetro da normatividade ou non dos lemas e das formas. Por que, ademais, das formas? Basicamente, porque a normatividade do lema non implica a normatividade da forma. Por exemplo, *fosemos*, forma totalmente correcta ata xullo de 2003 en que entrou en vigor a actual normativa, remítese ao lema "ser" ou "ir", normativos, sen embargo, *per se*, é un *token* que responde negativamente á normativa actual.

A economía lingüística e a reiteración impóñennos a obriga de aplicar este parámetro a algunhas desinencias xa que caracterizan formas dialectais recorrentes nos textos. É bastante frecuente, por poñer un exemplo, para a 3ª persoa do pretérito de indicativo dos verbos da 2ª conxugación, a aparición de formas coa vogal temática "i" no sitio da normativa "e". Coa

4. Conxuntos de desinencias para as que se proporcionan uns valores. Por exemplo o grupo G1 consta de:
   o: masculino singular
   a: feminino singular
   os: masculino plural
   as: feminino plural
   Deste xeito, para todos os substantivos e adxectivos que rematan en "-o", fan o feminino en "-a" e forman o plural engadindo "-s", tipo "neno", introducimos só o lema e a raíz correspondente, sen necesidade de ter que incluír completas as formas masculina e feminina cos respectivos plurais.

estrutura que empregamos en XIADA a opción máis rendible é a de caracterizar "-iu" no seu valor de "ei3s0"[5] como non normativa no grupo V2 segundo o que se conxugan todos os verbos regulares da 2ª conxugación.

Outro exemplo: os grupos G2 e G3 foron creados para flexionar os substantivos e adxectivos cuxo masculino singular remataba en *-án* e facían o feminino en *-á* e *-ana* respectivamente. A aplicación da normativa oficial reduce considerablemente o número de elementos que van ao grupo G3 incrementando o G2[6]. Ao igual que sucedía cos lemas, debemos manter as características dos dous grupos para que o analizador recoñeza e caracterice as formas existentes nos textos anteriores á normativa oficial actual pero tamén temos que introducir as modificacións precisas para que as novas formas normativas sexan recoñecidas e analizadas. A opción máis intelixente parece ser a caracterización das desinencias *-ana* no G2 e *-á* no G3, cos respectivos plurais, como non normativas.

É interesante e sobre todo útil, desde o punto de vista lingüístico, diferenciar na estrutura do lexicón módulos segundo o tipo de léxico que conteñan e con que poidan ser combinados. Para iso dispuxemos no deseño da estrutura, e obviamente implementamos na construción do lexicón, dun campo máis no que se caracteriza o lema como pertencente ao léxico común ou ao técnico-científico, especificando neste último o ámbito no que se clasifica: administración, economía, medicina, etc.

Relacionado co punto anterior está a inclusión na estrutura do lexicón dun indicador que nos marca a procedencia de cada un dos lemas. Nos textos actuais preséntanse formas que non aparecen nos dicionarios-vocabularios preceptivos da lingua galega –*DRAG* (1997) e o máis recente *VOLG* (2004)–, ben porque son termos técnicos para os que se acaba de propoñer unha denominación, ben porque se documentan cun uso categorial distinto. Non obstante, a súa introdución no lexicón é imprescindible para o recoñecemento e caracterización dos textos. Coa indicación da

5. Pretérito de Indicativo, 3º persoa do singular, non aplica xénero.

6. Na normativa oficial anterior seguían o esquema *-án / -ana* (as raíces e lemas destas formas remitíanse en XIADA ao grupo de derivación G3), mentres que na actual se acomodan ao esquema *-án / -á* (debemos remitir as raíces e lemas ao grupo de derivación G2) as formas seguintes: *afgán, alazán, alemán, barregán, bosquimán, capitán, catalán, ermitán, escribán, gardián, musulmán, rufián, sancristán, sultán e truán*.
O G2 consta das seguintes terminacións, que caracterizamos formalmente como:
án: masculino singular
áns: masculino plural
á: feminino singular
ás: feminino plural
ana: feminino singular (*)
anas: feminino plural (*)

fonte achegamos datos concretos sobre a localización do lema e a fiabilidade da forma.

Finalmente, polo que respecta aos parámetros de tipo de léxico e fonte, consideraremos que o *VOLG* (2004), como texto sancionador máis recente e posuidor dun número de entradas aceptable, definirá os lemas pertencentes ao léxico común ou principal do lexicón de XIADA.

Para que a estrutura do lexicón sexa permeable e máis accesible á identificación e asignación dos distintos modelos formais que o integran, inserimos no deseño da estrutura un *representante de grupo* que serve de modelo para cada un dos grupos existentes en XIADA. Así, o *representante de grupo* do grupo derivacional G1 é *neno*.

Recapitulando, o lexicón de XIADA, ademais de conter a información morfolóxica que permite a identificación e caracterización morfolóxica plena de calquera palabra galega, responde aos parámetros seguintes: indicación sobre a normativa oficial, ámbito léxico e procedencia da forma.

Unha vez deseñada a estrutura do lexicón, implementámolo coas entradas do *VOLG* (2004) e coas 70.000 formas máis frecuentes de *CORGA*, o que vén sendo algo máis de 51.000 lemas. Na seguinte táboa podemos ver o número de lemas e raíces das categorías principais do lemario de Xiada:

CLASE GRAMATICAL	LEMAS	RAÍCES	FORMAS
Adverbio	1.906	1.916	1.916
Adxectivo	13.574	13.625	44.899
Substantivo	28.806	28.943	62.074
Verbo	6.690	9.113	367.220

Un lexicón non ten por que estar destinado unicamente a etiquetar e lematizar un texto. Se é o suficientemente detallado e complexo pode servir de punto de partida para outras utilidades como a construción dun corrector ortográfico ou un conxugador de verbos.

3. PREPROCESADOR

Lingüisticamente, unha vez elaborados os modelos formais que van acoller as raíces e grupos de derivación e implementado o lexicón computacional, estamos en disposición de etiquetar automaticamente calquera texto galego contemporáneo. Trátase "simplemente" de executar o

programa que o equipo informático do proxecto está desenvolvendo para traballar con arquivos codificados en XML (cfr. http://corpus.cirp.es/xiada), porque ese é o formato no que están os textos de *CORGA*.

Un preprocesador, en xeral, delimita as distintas unidades léxicas que hai nun texto. Pode parecer sinxelo pero non é tarefa banal segmentar formas verbais con pronomes enclíticos ou separar os formantes dunha contracción. Seguimos traballando para que cometa menos erros pero é inevitable que nalgúns casos de ambigüidades falle de vez en cando. É un traballo informático pero no que o equipo lingüístico participa coa construción de regras que gobernan o comportamento do preprocesador. Por exemplo, cando estabamos comprobando o funcionamento das formas verbais con clíticos, hai xa bastante tempo, atopámonos con que para o *token* "semana" propoñía unha etiqueta na que o segmentaba en "se" (raíz verbal), "ma" (contracción dos pronomes *me* máis *a*) e o último constituínte formábao o alomorfo do pronome acusativo de terceira persoa *na*. Vimos entón a necesidade de construír unha regra que impedise análises coma esta, polo que creamos unha na que indicabamos que na mesma secuencia non pode haber dous clíticos de acusativo e, ademais, *no*, *na*, *nos*, *nas*, pronome átono acusativo de terceira persoa, só será unha etiqueta posible se vai enclítico a unha forma verbal rematada en ditongo e en ningún outro caso.

O preprocesador, no noso caso, ademais, despois de delimitar as unidades, asígnalle a cada unha os lemas e etiquetas posibles.

Así, a saída do documento preprocesado proporcionaranos o texto analizado contendo para cada palabra gráfica ou *token* todas as posibles análises morfolóxicas.

Por exemplo, o preprocesador ante un *token* como "polos", no que non só hai ambigüidade morfolóxica senón que tamén a segmentación da forma en constituíntes varía, proporciona as seguintes análises:

(1) Unha alternativa na que o clasifica como substantivo común masculino plural cuxo lema é *polo*:

```
<alternativa>
        <constituínte>
                <forma>polos</forma>
                <etq_lema>
                        <etiqueta>Scmp</etiqueta>
                        <lema>polo</lema>
                </etq_lema>
        </constituínte>
</alternativa>
```

(2) Noutra alternativa segméntao en dous constituíntes definindo o primeiro como a preposición *por* e o segundo como o artigo determinado masculino plural cuxo lema é *o*:

```
<alternativa>
        <constituínte>
                <forma>por</forma>
                <eta_lema>
                        <etiqueta>P</etiqueta>
                        <lema>por</lema>
                </eta_lema>
        </constituínte>

        <constituínte>
                <forma>os</forma>
                <eta_lema>
                        <etiqueta>Ddmp</etiqueta>
                        <lema>o</lema>
                </eta_lema>
        </constituínte>
</alternativa>
```

(3) Na última alternativa segméntao en *pos*, 2ª persoa do presente de indicativo dos verbos *pór* e *poñer* e mais o pronome átono de acusativo de terceira persoa masculino plural cuxo lema é o:

```
<alternativa>
        <constituínte>
                <forma>pos</forma>
                <eta_lema>
                        <etiqueta>Vpi2s0</etiqueta>
                        <lema>poñer</lema>
                </eta_lema>

                <eta_lema>
                        <etiqueta>Vpi2s0</etiqueta>
                        <lema>pór</lema>
                </eta_lema>
        </constituínte>

        <constituínte>
                <forma>os</forma>
                <eta_lema>
                        <etiqueta>Raa3mp</etiqueta>
                        <lema>o</lema>
                </eta_lema>
        </constituínte>
</alternativa>
```

Neste punto teriamos o documento coas etiquetas e lemas posibles. Dado que para poder facer as buscas nel necesitamos ter a etiqueta correcta da palabra e non todas as posibles, temos dúas opcións:

(1) Determinar manualmente cal é a correcta.
(2) Que un sistema automático decida cal é a correcta.

É aquí onde entra en xogo o punto seguinte, o etiquetador, que "simplemente" escolle a etiqueta e o lema correctos para cada unha das palabras.

4. ETIQUETADOR

Antes de que se poida empregar un etiquetador automático estatístico, como é o caso, é necesario adestralo desambiguando manualmente un subconxunto de textos. Obviamente canto maior sexa o volume de textos que se desambigüen á man máis probabilidades hai de que acerte o etiquetador. O noso obxectivo, ademais, é que, fronte ao 95/96% de acerto que presentan os etiquetadores existentes para o inglés –para o galego non hai estatísticas–, o noso etiquetador alcance un acerto do 99%.

A utilización do preprocesador e do etiquetador automático vainos resultar útil para atender dúas frontes: por unha banda, revisar as análises obtidas automaticamente e, pola outra, tomar nota das formas descoñecidas para mellorar o seu funcionamento e incrementar o lexicón computacional.

Esperamos que proximamente sexa posible consultar tanto o lematizador e o etiquetador en rede como un subcorpus de *CORGA* lematizado e etiquetado.