O macroproxecto CORGA-XIADA: corpus, etiquetador e flexionador

Eva Domínguez, Mario Barcala, Marisol López, Guillermo Rojo

Introdución

Cando se fala de humanidades dixitais e tecnoloxías lingüísticas para a investigación en humanidades emprendidas en Galicia e en galego é ineludible adicar un apartado, por unha banda, ao Corpus de Referencia do Galego Actual (CORGA) (http://corpus.cirp.gal/corga) e, pola outra, ao sistema XIADA (http://corpus.cirp.gal/xiada), isto é, a aquelas ferramentas e recursos relacionados coa súa anotación morfosintáctica: o etiquetador morfolóxico, o corpus de adestramento, a base de datos léxica e, máis recentemente, o flexionador e o dicionario de frecuencias. Deles, os 3 primeiros esíxeos a anotación automática do corpus cun etiquetador de tipo probabilístico, mentres que os dous últimos poden considerarse produtos derivados.

Neste póster descríbese brevemente a última versión do CORGA: características esenciais, posibilidades que ofrece a aplicación de consulta para a recuperación de información e a súa anotación morfosintáctica. Así mesmo, dáse a coñecer a versión pública do etiquetador e preséntase un flexionador verbal e nominal conectado co CORGA.

O corpus

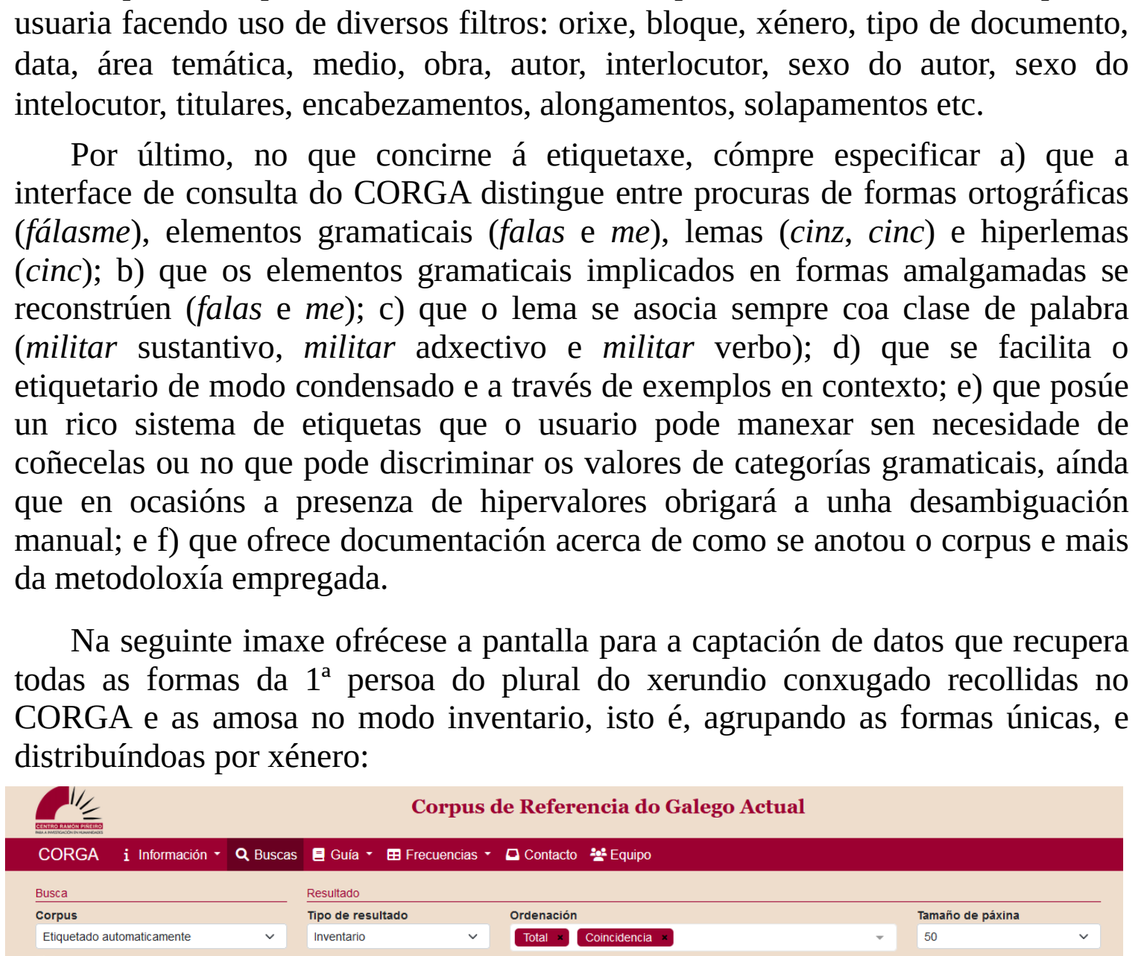
O CORGA é un corpus documental aberto, de acceso libre e de balde, que abrangue cronoloxicamente dende 1975 ata a actualidade e que se considera representativo da lingua galega. Contén na súa versión 4.1 de 2024 **45.660.649 palabras** (54.737.277 elementos gramaticais), das que case 700.000 pertencen a mostras do rexistro oral. O corpus está codificado no estándar XML e anotado automaticamente para facilitar a recuperación de información.

A aplicación de consulta permite realizar procuras mediante formas ortográficas ou información gramatical, tendo en conta a contigüidade ou a proximidade entre elementos, con ou sen operadores booleanos e metacaracteres, e amosa os resultados en forma de frecuencias (táboas ou gráficas), concordancias (a expresión obxecto da busca centrada e cun pequeño contexto anterior e posterior) e inventario (formas únicas da expresión buscada distribuídas ademais por frecuencias e calquera outro dos parámetros de clasificación).

A aplicación permite ademais crear subcorpus virtuais á medida da persoa usuaria facendo uso de diversos filtros: orixe, bloque, xénero, tipo de documento, data, área temática, medio, obra, autor, interlocutor, sexo do autor, sexo do intelocutor, titulares, encabezamentos, alongamentos, solapamentos etc.

Por último, no que concirne á etiquetaxe, cómpre especificar a) que a interface de consulta do CORGA distingue entre procuras de formas ortográficas (*fálasme*), elementos gramaticais (*falas* e *me*), lemas (*cinz*, *cinc*) e hiperlemas (*cinc*); b) que os elementos gramaticais implicados en formas amalgamadas se reconstrúen (*falas* e *me*); c) que o lema se asocia sempre coa clase de palabra (*militar* sustantivo, *militar* adxectivo e *militar* verbo); d) que se facilita o etiquetario de modo condensado e a través de exemplos en contexto; e) que posúe un rico sistema de etiquetas que o usuario pode manexar sen necesidade de coñecelas ou no que pode discriminar os valores de categorías gramaticais, aínda que en ocasións a presenza de hipervalores obrigará a unha desambiguación manual; e f) que ofrece documentación acerca de como se anotou o corpus e mais da metodoloxía empregada.

Na seguinte imaxe ofrécese a pantalla para a captación de datos que recupera todas as formas da 1ª persoa do plural do xerundio conxugado recollidas no CORGA e as amosa no modo inventario, isto é, agrupando as formas únicas, e distribuíndoas por xénero:



O etiquetador

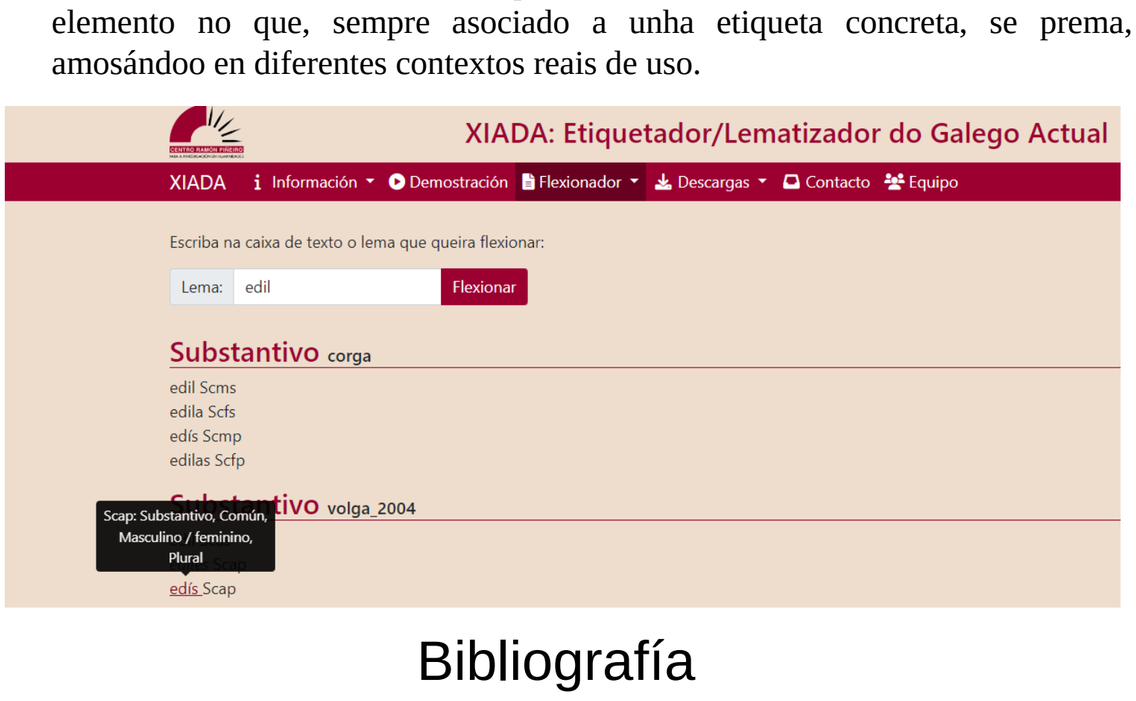
O Etiquetador/Lematizador do Galego Actual (XIADA) (http://corpus.cirp.gal/xiada/) é un etiquetador fundamentalmente estatístico baseado nos modelos de Markov de grao 2, é dicir, modelos estatísticos nos que os cálculos relativos ás etiquetas morfosintácticas dunha palabra dependen das dúas anteriores (Brants 2000). Foi desenvolvido conxuntamente entre o Centro Ramón Piñeiro para a investigación en humanidades e o grupo COLE das universidades da Coruña e Vigo para etiquetar automaticamente o CORGA e contribuír ao deseño e desenvolvemento de recursos que axuden á incorporación do galego ás Tecnoloxías da Información e a Comunicación e, ao tempo, posibilitar a obtención de datos no corpus para o estudo de aspectos léxicos, gramaticais, discursivos etc.

O etiquetador contén, entre outros, módulos de segmentación de oracións, tokenización, anotación morfosintáctica e lematización. Conta ademais con regras lingüísticas de restrición (Barcala *et al.*, 2007) e de poda (Domínguez, 2016) que contribúen a incrementar sensiblemente a súa taxa de acerto. En concreto, en relación á súa cobertura, precisión e robustez na análise, tendo en conta os datos proporcionados por Domínguez, Barcala e Molinero (2009), pode afirmarse que a súa precisión está en liña co estado da arte na materia (96%). O código do etiquetador foi liberado no ano 2019, xunto cos recursos que utiliza (https://github.com/crpih/xiada), mais agora ponse á disposición da sociedade o seu emprego para etiquetar arquivos de texto, de xeito que calquera persoa poida anotar o seu corpus sen necesidade de coñecementos técnicos. Velaquí unha mostra na que destacamos a anotación de formas con seseo:

O flexionador

A ausencia dun flexionador verbal e nominal integrado da lingua galega e mais a existencia da base de datos léxica de XIADA leváronnos a desenvolver un flexionador cuxos principios se poden sintetizar nos seguintes termos:

- Susténtase na base de datos léxica de XIADA.
- Ofrece a flexión de calquera lema contido no lexicón. Para as clases de palabras variables amosa flexión de xénero, número, modo, tempo, persoa etc. No caso das invariables só contén as propias formas e mais pode acoller outras elativas ou apreciativas (*cerca*, *cerquiña*, *cerquísima*).
- O lema está sempre asociado a unha clase de palabra (*sobre*-S, *sobre*-P).
- Ofrécense tantos paradigmas por lema como modelos flexivos diferentes este presente (na figura, *edil/edís* vs. *edil/edila/edís/edilas*).
- Indícase a normatividade de todas as formas: a redonda sinala que se adecúa ás NOMIG actuais, mentres que a itálica sinala que as infrinxe).
- Conéctase co CORGA, de xeito que se accede directamente ás ocorrencias do elemento no que, sempre asociado a unha etiqueta concreta, se prema, amosándoo en diferentes contextos reais de uso.



Bibliografía

Barcala, Fco. Mario / Molinero, Miguel A. / Domínguez, Eva. 2007. XML Rules for Enclitic Segmentation. En Roberto Moreno-Díaz *et al.* (eds.), *Computer Aided Systems Theory*, 4739, *Lecture Notes in Computer Science*. Berlin-Heidelberg-New York: Springer-Verlag, 273- 281.

Brants, Thorsten. 2000. TnT - A Statistical Part-of-Speech Tagger. En *Proceedings of the Sixth Applied Natural Language Processing Conference*. Seattle-Washington: Association for Computational Linguistics, 224-231.

Domínguez, Eva Mª. / Barcala, Fco. Mario / Molinero Miguel A. 2009. Avaliación dun etiquetador automático estatístico para o galego actual: Xiada. *Cadernos de Lingua* 30-31, 151-193.

Domínguez, Eva Mª. 2016. O etiquetador probabilístico de XIADA e o seu teito de acerto: a elaboración de regras lingüísticas. En Manuel González (ed.), *Lingua, pobo e terra. Estudos en homenaxe a Xesús Ferro Ruibal*. Santiago de Compostela: Xunta de Galicia - Centro Ramón Piñeiro para a investigación en humanidades, 213-232.

printed by MegaPrint Inc. www.postersession.com