De lexicografía galega. Un diccionario máquina

ANTÓN SANTAMARINA

No ILG e na RAG levamos algún tempo traballando nun proxecto de lexicografía galega con tratamento informático dos datos. Dunha maneira moi breve podemos dicir que o primeiro paso, donde sairán varios posibles diccionarios, consiste na entrada de textos de tódolos rexistros do galego moderno; no seu tratamento cun programa de concordancias e lematizacións; e na almacenaxe e integración nunha grande base de datos.

Dous dos momentos principais desta operación, a introducción de datos e o seu manexo na fase de lematización, son coñecidos porque contan cos seus manuais respectivos (aínda que de uso interno). Pero hai un instrumento intermedio que, por non ser usado polos operadores, non necesita coñecementos para o seu manexo e non quedará descrito en ningún manual. Por iso coidamos de interés dar unha noticia del.

Para automatizar ata onde sexa posible o proceso de lematiza-las palabras de cada obra concordada (ou conxuntamente de varias obras) preparouse un *Diccionario máquina* (*DM*) no que están flexionadas tódalas palabras da versión provisional (corrixida) do *Vocabulario ortográfico da lingua galega* (*VOLGA*) (1). Está claro que non é un diccionario que teña unha finalidade parecida a ningún dos tipos de diccionarios que se fabrican (por iso lle chamamos "diccionario máquina"); nin é tampouco un diccionario no que poidan aparecer tódalas formas usadas polos autores do corpus, pois está baseado no *VOLGA*, que só dá entrada a unhas 45.000 palabras, escollidas cun criterio purista.

Se se examina a edición provisional do *VOLGA* pódese comprobar que no campo de "categoría" hai bastante información gramatical (*t*ransitivo, *i*ntransitivo, *p*ronominal, *a*dxectivo, *s*ustantivo, *s*ustantivo e *a*dxectivo, etc.). Para o *DM* non se precisa tanta información e ademais a casuística a que nos podía levar facer un modelo flexivo para cada categoría conduciríanos a unha serie de distincións imposibles de controlar tanto por parte dos constructores do *DM* como, despois, polos lematizadores; por esta razón foron reducidas e unificadas. Nomeadamente ás seguintes clases de palabras:

(1) Instituto da Lingua Galega e Real Academia Galega, *Vocabulario ortográfico da lingua galega (versión provisional)*, Santiago, 1989, edición non venal.

1 artigo
2 contracción de prep + artigo
3 sustantivo
4 adxectivo
5 adxectivo/sustantivo (*lugués*)
6 pronome persoal
7 contracción de preposición + pronome persoal
8 demostrativo
9 contracción de preposición + demostrativo
10 posesivo
11 relativo
12 interrogativo
13 indefinido
14 contracción de preposición + indefinido
15 numeral
16 verbo
17 adverbio
18 preposición
19 conxunción
20 interxección
21 locución (de varios tipos).

Nun diccionario máquina, no campo "categoría" non pode aparecer, por exemplo, "(verbo) primeira persoa de plural do presente de indicativo"; nin tampouco unha abreviatura facilmente memorizable. Por esta razón reducíronse as categorías a unha cifra de tres díxitos (que ocupa pouca memoria e é facilmente traducida a unha táboa de equivalencias en "palabras"); así: do 001 a 018 corresponde ó artigo, do 021 a 038 á contracción de preposición e artigo, e así sucesivamente.

A secuencia dos números é a habitual nos paradigmas gramaticais: masculino singular - feminino singular - etc. 061 *cativo*, 063 *cativa*, 065 *cativos*, 067 *cativas*, 069 *cativiño*, 071 *cativiña*, 073 *cativiños*, 075 *cativiñas*; ou no verbo: infinitivo - xerundio - participio - presente de indicativo - imperfecto etc. e dentro de cada tempo, singular 1ª persoa, 2ª etc. E así con cada unha das partes flexionais. Deste modo, cando se lematice o vocabulario dun autor, ou todo o vocabulario do corpus, baixo de cada lema virán ordenadas as distintas formas paradigmaticamente e non alfabeticamente; é dicir, aparecerá (501) *ser*, (503) *seres*, (507) *serdes*, (509) *seren*, (511) *sendo*, (513) *sido*, (525) *son*, (527) *es*, (529) *é*... (537) *era*... (549) *fun*,... (561) *fora*... e non *é, era, es, fora, fun, sendo, ser, serdes, seren, seres, sido, son*.

A elaboración dun *DM* para unha lingua de flexións complexas como é o galego non se pode facer manualmente máis ca nuns poucos casos; levaría unha enormidade de tempo e daría lugar a unha cantidade grande de erros. Non vale a pena pormenorizar como se fixo. Daremos só unha breve noticia diso. O material do *VOLGA* foi sendo librado por clases de palabras para ficheiros individuais e despois cada ficheiro tratado independentemente, ás veces con fragmantacións

parciais. Ás clases invariables (adverbios, preposicións, etc.) atribuíuselle-la súas categoría, impar (normativa) ou par (non normativa), segundo os casos. Para as partes variables (menos o artigo, as contracciós e o pronome) construíronse modelos de flexión e repasouse a base de datos do *VOLGA* atribuíndo a cada palabra o seu modelo mediante unha clave. Despois varios programas fixeron as flexións de maneira automática. Para os artigos, contraccións e pronomes e para as dúas ducias de verbos irregulares fíxose a flexión de modo manual.

Díxose antes que o *DM* estaba baseado no *VOLGA*, e por tanto estaban excluídas del as palabras non normativas. De toda maneira, nunha lingua como a galega no período que nos ocupa, en estado de constitución, hai moita variante. Un diccionario máquina que non considerase estas variantes non admitidas no *VOLGA* vería seriamente minguada a súa eficacia como máquina. Realmente, xa no mesmo *VOLGA* se intruducen bastantes palabras de uso moi frecuente precedidas por un asterisco para indicar que non se lles dá rango de normativas. Pero ademais destas entradas, na flexión das palabras variables, engadíronse ás formas flexivas normativas, tamén as formas non normativas máis frecuentes. Dito cun exemplo, o paradigma de *canción* no *DM* será

043	S_fs	canción
047	S_fp	cancións
048	S_fp*	cancións
048	S_fp*	cancíois.

Se queremos un caso de complexidade moito maior, velaquí o paradigma do perfecto de facer

549	PrfI1	fixen
549	PrfI1	fíxe-n
550	PrfI1*	ficen
550	PrfI1*	fíce-n
550	PrfI1*	figuen
550	PrfI1*	fígue-n
551	PrfI2	fixeches
551	PrfI2	fixéche-l
552	PrfI2*	fixeche
552	PrfI2*	ficeches
552	PrfI2*	ficéche-l
552	PrfI2*	ficeche
552	PrfI2*	figueches
552	PrfI2*	figuéche-l
552	PrfI2*	figueche
553	PrfI3	fixo
554	PrfI3*	fizo
554	PrfI3	figo
554	PrfI3*	fezo
555	PrfI4	fixemos
555	PrfI4	fixémo-l
556	PrfI4*	ficemos
556	PrfI4*	ficémo-l
556	PrfI4*	figuemos
556	PrfI4*	figuémo-l
557	PrfI5	fixestes
557	PrfI5	fixéste-l

558	PrfI5*	fixéstedes
558	PrfI5*	fixéchedes
558	PrfI5*	ficéstede-l
558	PrfI5*	ficéchede-l
558	PrfI5*	ficestes
558	PrfI5*	ficéste-l
558	PrfI5*	figuestes
558	PrfI5*	figuéste-l
558	PerI5*	figuestedes
558	PrfI5*	figuéstede-l
558	PrfI5*	figuechedes
558	PrfI5*	figuéchede-l
559	PrfI6	fixeron
559	PrfI6	fixéro-n
560	PrfP6*	ficeron
560	PrfP6*	ficéro-n
560	PrfI6*	figueron
560	PrfI6*	figuéro-n

(Nas formas con guión, o guión pertence á desinencia verbal [suplindo o *-r*, o *-s* ou o *-n* elididos]; o *l* e mailo *n* entran nun campo chamado "enclítico" que nos axuda a diferenciar formas como ***cánta-los*** e ***cánta-nos;*** á súa vez, ambas distintas de ***canta*** *nos*. Ver máis abaixo.).

Sería metodoloxicamente un erro nesta mestura de formas prescindir do dato de se unha forma é ou non normativa; é dicir, se está de acordo co *VOLGA* e coas *Normas ortográficas e morfolóxicas* (2). Iso sinalouse usando as cifras impares para as formas normativas e as pares para as outras. Na categoría expresada en abreviatura (PrfI1 = Perfecto de Indicativo persoa 1 etc.) indícase mediante asterisco o carácter non normativo dunha forma (550 PrfI1*). Indicar se unha forma é non normativa mediante cifra par e asterisco é unha redundancia que pode ter algunha utilidade. Por suposto, as palabras que no *VOLGA* aparecen con asterisco levan todas números pares no *DM*. Desta maneira o diccionario máquina tén información potencial que vai máis alá do seu uso inmediato de servir para lematizar; unha das posibles utilidades ulteriores sería a de servir de material bruto para un programa corrector de textos (se é que algunha compañía informática chega a interesarse por iso); abondaría con quitar del as formas pares. Podería servir para extraer un diccionario de verbos conxugados. E quén sabe que máis.

Se as formas oblicuas non normativas dunha palabra se colocan despois das formas normativas, isto mesmo ocorre tamén coas variantes rectas das formas non normativas; desta maneira no *DM* teremos agrupadas despois da forma normativa dunha entrada tódalas súas variantes fonéticas, é dicir, ó ládo de *ciruxián* virán (aquí alfabeticamente) *ceruxano*, *ciruxano, surxán, zuruxano* e *zurxán*. Así poderemos ter agrupadas en boa medida formas que na simple concordancia virían moi dispersas por todo o alfabeto.

Aínda que o *DM* ten como principio mante-la distinción entre normativo e non normativo, hai lugares en que o respecto da norma levaría a un atranco insuperable. É sabido que a ortografía do galego estableceu como norma soldar di-

(2) Instituto da Lingua Galega e Real Academia Galega, *Normas ortográficas e morfolóxicas da lingua galega*, Santiago, 1982.

rectamente o pronome enclítico ó verbo e alterar graficamente a acentuación. Os programas informáticos operan con "palabras" e só a custo dunha programación complicadísima sería posible que formas como *díxome, dixéronnolo* ou *dixémoschellelo* fosen concordadas como dúas palabras (ou tres ou catro, respectivamente); iso obrigounos a altera-los textos na fase de preedición separando por guión o verbo do pronome e deixando a acentuación do verbo coma se fose forma exenta; cando amais de énclise hai asimilación da última consoante ó pronome ou ó artigo tamén se fai a separación pero alterando a acentuación da forma verbal polas razóns que explicamos no manual do editor. A alternativa era ou sacrifica-la automatización ou sacrifica-la ortografía; pareceunos máis oportuno o segundo. A fin de contas o que resulte despois da expurga dos textos do corpus será o material bruto para a redacción do diccionario (ou dos diccionarios); e en calquera caso sempre se respectou o contido fonético e morfolóxico do texto aínda que se alterase, mínimamente, a súa representación gráfica.

Nun diccionario deste tipo é moi grave o problema da homonimia (porque non hai definicións). Isto está resolto nos casos en que as palabras homónimas pertencen a clases diferentes porque quedan discriminadas de modo automático; nos outros casos a discriminación houbo que facela de modo manual extraendo por categorías tódalas palabras que tiñan lemas coincidentes para engadirlles unha nota referida ó seu significado. Hai amais diso as homonimias que xorden dentro do propio paradigma (*cantar* pode ser varias cousas distintas: infinitivo invariable, infinitivo 1ª persoa, infinitivo 3ª persoa, futuro de subxuntivo 1ª persoa e futuro de subuntivo 3ª persoa). Estas resolvéronse de modo automático na numeración e abreviatura das categorías. Velaí un exemplo de complexidade simple:

PALABRA		LEMA	CATEGORIA
cacho	1	cachar = coller	525 PrsI1
	2	cachar = roturar	525 PrsI1
	3	cacho	061 AX_ms
	4	cacho (recipiente)	041 S_mas
	5	cacho 'pedazo'	041 S_ms
	6	cacho 'acio'	041 S_ms

(*PrsI1* = Presente de indicativo persoa 1; *AX_ms* = Adxectivo masculino singular; *S_ms* = Sustantivo masculino singular)

Digamos aquí, de paso, que o traballo fundamental do lematizador consistirá en escoller para a palabra *cacho* unha das 6 opcións (ou unha que hai sempre en branco para casos non previstos) segundo o que signifique no contexto que mostrará a pantalla simultaneamente.

Este diccionario máquina non é unha obra pechada. Amais das formas de partida (as do *VOLGA* con flexións normativas e non normativas) irán incorporándoselle as variantes ou formas novas de cada autor a medida que se vaian lematizando. Porque está previsto que na fase de lematización as palabras novas e variantes se flexionen e se incorporen ó *DM*. Castelao en *Cousas* usa palabras como *abrouxa, acarrexara, adeprendeu,* etc., que non figuran no *VOLGA* porque como normativas só se propoñen *abouxar, carrexar, aprender*. Nunha das fases da lematización aquelas formas serán reducidas á "non marcada", flexionadas, atri-

buídas a un lema, e a súas variantes flexivas intercaladas sempre como formas non normativas entre as formas de *abouxar, carrexar* e *aprender.* Con isto está claro que a eficacia do *DM* será maior a medida que se avance no proceso de lematización porque cando no sucesivo volvan a aparecer *abrouxar, acarrexar* etc., en calquera tempo ou persoa que estean flexionadas, serán automaticamente atribuídas a *abouxar, carrexar,* etc.

Para realizar todo este traballo usáronse ordenadores persoais. A razón é que non se dispuxo no seu momento dun ordenador grande e por outra parte todo o traballo de preedición das obras resultaba máis cómodo cun programa de tratamento de textos dos dispoñibles no mercado, con posibilidades de usar diacríticos sen necesidade de máis manipulación. Son necesarios de toda maneira PCs de certa potencia porque o *DM* no momento da partida tén por riba do millón trescentas mil formas (ocupa mais de 50MB); e a medida que se lematicen obras aumentará.

Este *DM* vai estar nun "servidor" de 12MB de RAM e 314MB de ROM; conectados a el, en rede local, estarán cinco PCs de 60MB; os seis operadores que se prevé que traballen simultaneamente terán desde o seu posto acceso o *DM* para facer todo o traballo de lematización. Despois de cada obra lematizada incorporaráselle o *DM* o diccionario parcial xerado.