A nomenclatura científica binomial no *Corpus de Referencia do Galego Actual*: unha proposta de análise extensible a outros corpus e linguas

Eva Mª Domínguez Noya / Vítor Míguez Rego
VI Congreso internacional de lingua, lingüística e tecnoloxía (techLING2021-UVigo-T&P)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
INSTITUTO DA LINGUA GALEGA (ILG)
CENTRO RAMÓN PIÑEIRO PARA A INVESTIGACIÓN EN HUMANIDADES

Introdución

A nomenclatura científica permite a identificación exacta de calquera organismo, vivo ou extinto, e resolve os problemas de sinonimia das denominacións vernáculas. Deste xeito *polbo*, *pulpo*, *pop*, *polvo* etc. identifícanse internacionalmente mediante *Octopus vulgaris*. Nos últimos anos a detección e extracción de termos binomiais espertou un interese crecente no ámbito do recoñecemento de entidades nomeadas, especialmente na minería de textos da literatura biomédica [1] ou na busca de correspondencias coa denominación ordinaria das especies biolóxicas [2]. O noso interese, non obstante, céntrase na súa anotación morfosintáctica automática no *Corpus de Referencia do Galego Actual* (CORGA) [3], categorización potencialmente relevante para outros corpus, dado que o sistema de nomenclatura binominal, ademais de identificar unha especie inequivocamente, caracterízase a efectos lingüísticos por conformar expresións latinas ou latinizadas, o cal as afasta do acervo léxico dunha lingua concreta, coma o galego, e converte esta proposta nunha solución extrapolable a outras linguas.

A convención que determina escribi-la inicial do xénero en maiúscula conduce no CORGA a unha análise desagregada errónea da expresión binomial na que a forma correspondente ó xénero se caracteriza como nome propio. A continuación trataremos de dar conta do problema, revisa-la súa caracterización noutros corpus e ofrecer unha proposta de análise.

Análise do problema

O CORGA é un corpus documental aberto representativo da lingua galega actual, codificado en XML e enriquecido con anotación morfosintáctica mediante a ferramenta XIADA [4], un etiquetador de tipo estatístico. Na súa primeira capa de análise lingüística [5] non tivemos en conta as entidades. A identificación destas como substantivos propios no interior de secuencia obedece, *grosso modo*, á presenza dunha maiúscula inicial seguida de caracteres en minúscula, de forma que toda palabra que comeza con maiúscula se identifica como substantivo propio. E é precisamente esta caracterización da parte relativa ó xénero da denominación científica, xunto coa clasificación como substantivo ou adxectivo do subtipo posterior, a xeito de unidades independentes, a que orixina que examinémo-las expresións binomiais da nomenclatura científica, co fin de favorece-la súa anotación conxunta e corrixi-las malogradas análises actuais. Trátase pois de integrar estas formas de modo coherente coas características dun recurso que conta xa con 40 millóns de palabras.

A investigación inicial sobre as denominacións científicas binomiais ou trinomiais –en que consisten, cales son as súas características etc.– lévanos a consideralas expresións latinizadas que conforman unha unidade multipalabra cuxo comportamento é similar ó dun substantivo, polo que propoñémo-la súa integración nesta clase de palabras. Non obstante, rexeitamos tratalos como propios por entender que tanto a denominación vernácula como a científica son comúns, inda que por convención a inicial do xénero se escriba con maiúscula, feito que conduce na anotación automática á interpretación xeneralizada como propio.

Pola súa parte, a consulta a outros corpus anotados, de referencia ou técnicos, do galego, español e catalán presenta resultados heteroxéneos, como demostra a selección recollida na táboa seguinte:

Termo binomial	**TILG** [6]	**CTAG** [7]	**CORPES** [8]	**CTILC** [9]	**CT** [10]
Quercus robur; *Quercus ilex** (carballo; aciñeira)	2 unidades: Palabra estranxeira + Palabra estranxeira	1 unidade: NCMS0 (substantivo común masc. sg.)	2 unidades: Substantivo propio + estranxeirismo	1 unidade: NC (non codificado)	*1 unidade: N5-66 (nome común, xénero e número pendentes)
Castanea sativa (castiñeiro)	2 unidades: Palabra estranxeira + Palabra estranxeira	1 unidade: NCFS0 (substantivo común fem. sg.)	2 unidades: Substantivo propio/ Estranxeirismo + adxectivo fem. sg. (lema *sativo*)	1 unidade: NC	1 unidade: N5-66
Sus scrofa (xabaril)	2 unidades: Palabra estranxeira + Palabra estranxeira	1 unidade: NCFS0	2 unidades: posesivo, masc. pl. (lema *suyo*) + ?? (descoñecido)	1 unidade: NC	1 unidade: N5-66

Dos corpus examinados, entre os que tamén se inclén o BNC [11] e o CRPC [12], ningún reserva unha etiqueta específica para a nomenclatura científica e soamente o CTAG e mailos dous do catalán, CTILC e CT, consideran esas expresións unha unidade multipalabra. O CTAG e o CT etiquétanas ademais como nomes comúns, aplicando no caso do primeiro xénero e número, mentres que o segundo deixa indeterminados tales atributos. O resto dos corpus en xeral analiza os membros por separado e a súa categorización tende á que amosa o CORPES: propio se o xénero non coincide con ningunha palabra do léxico común e aparece coa inicial en maiúscula, e estranxeirismo para o subtipo.

Referencias

1. Pafilis, E., *et al*. (2013): ‘‘The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text’’. *PLoS ONE*, 8(6), e65390.
2. Seideh, M. A. F., Fehri, H., & Haddar, K. (2016):‘‘Recognition and Extraction of Latin Names of Plants for Matching Common Plant Named Entities’’. *Automatic Processing of Natural-Language Electronic Texts with NooJ*, 132–144.
3. Centro Ramón Piñeiro para a investigación en humanidades: *Corpus de Referencia do Galego Actual* [3.2]. <http://corpus.cirp.gal/corga/>
4. Centro Ramón Piñeiro para a investigación en humanidades: *Etiquetador/lematizador do galego actual* [2.7]. <http://corpus.cirp.gal/xiada>
5. Domínguez Noya, E. M. (2013): *Etiquetaxe e desambiguación automáticas en galego: o sistema XIADA*. Tese de doutoramento. Universidade de Santiago de Compostela. <http://hdl.handle.net/10347/9587>
6. TILG: *Tesouro informatizado da lingua galega*. <http://ilg.usc.gal/TILG/>
7. CTAG: *Corpus Técnico Anotado do Galego*. <http://sli.uvigo.es/CTAG/>
8. CORPES XXI: *Corpus del Español del Siglo XXI*. <http://www.rae.es>
9. CTILC: *Corpus textual informatitzat de la llengua catalana*. <https://ctilc.iec.cat>
10. CT: *Corpus Tècnic*. <https://www.upf.edu/es/web/iula/corpus>
11. BNC: *Britsh National Corpus XML edition* [3.3.8]. <https://cqpweb.lancs.ac.uk>
12. CRPC: *Corpus de Referencia do Português Contemporâneo*. <http://alfclul.clul.ul.pt/CQPweb/crpcfg16/>
13. Pyle, Richard L. (2016): ‘‘Towards a Global Names Architecture: The future of indexing scientific names’’. *ZooKeys*, 550, 261–281.
14. Rouco, M., *et al*. (2019): *Lista de las aves de España*. SEO/BirdLife. <https://seo.org/wp-content/uploads/2019/05/ListaAvesdeEspaña2019.pdf>
15. Rivers, M. (2019): *European Red List of trees*. <https://doi.org/10.2305/IUCN.CH.2019.ERL.1.en>
16. *BOE* núm. 143, de 15/06/2019. https://www.boe.es/buscar/doc.php?id=BOE-A-2019-9026

A implementación dunha nova etiqueta

Á luz do tratamento unitario que o CTAG, o CTILG e o CT aplican sobre as denominacións científicas binomiais, así como da súa caracterización, decidimos tratalas tamén no CORGA como substantivos, pero non como comúns, senón que optamos por crear un novo subtipo que denominamos ‘nomenclatura científica’, xa que, por un lado, estas expresións pertencen a unha lingua distinta do galego e, por outro, ó individualizar estes elementos facilitámo-la súa recuperación e extracción na plataforma de consulta. Ademais, seguindo o CT e o CTILC, decidimos non concretar valores para xénero nin número.

A fin de implementa-la nova etiqueta no sistema (‘Snaa’, equivalente a substantivo, nomenclatura científica, masculino/feminino, singular/plural) e para que o etiquetador poida asignala, en primeiro lugar introducimos no *kérnel* ou núcleo de XIADA, arquivo onde consta cada etiqueta analizada polo menos unha vez, a análise no formato‘forma\tetiqueta\tlema’ da seguinte secuencia extraída do CORGA:

> Neste nivel podemos ver especies de gran interese económico como: ourizos de mar (Paracentrotus lividus), percebes (Pollicipes pollicipes), nécoras (Necora puber) ou os polbos (Octopus vulgaris) etc.

A seguir revisámo-lo corpus de adestramento empregado polo etiquetador para detecta-la posible presenza dalgunha destas expresións e reetiquetalas en función dos novos parámetros, de modo que os recursos sexan consistentes. Desbotando a opción de revisar manualmente o corpus debido ó seu tamaño (617.492 palabras ortográficas), aproveitamos que está dispoñible para a súa consulta en liña e recuperamos tódolos substantivos multipalabra formados por polo menos dous elementos, tanto comúns coma propios. Na revisión detectamos entre os comúns *ginkgo biloba* (3), *piñeiro pinaster* (2) e *estafilococus aureus* (1), mentres que entre os propios localizamos *Caenorhabditis elegans* (1) e *Tursiops Truncatus* (1). Como se constata, esta relación non está exenta de controversia e exemplifica as dificultades que supón traballar con textos reais que a miúdo conteñen erros tipográficos (a-b), caracteres alleos ó latín (c-d) ou usos incorrectos das maiúsculas (f-g):

	Forma rexistrada	Nome científico
a	*Lucamus cervus*	*Lucanus cervus*
b	*Cupresus* sp.	*Cupressus* sp.
c	*Sardiña pilchardus*	*Sardina pilchardus*
d	*Nécora puber*	*Necora puber*
f	*canis lupus familiaris*	*Canis lupus familiaris*
g	*Tursiops Truncatus*	*Tursiops truncatus*

En todos estes casos establecemos como lema a propia expresión binominal documentada no corpus, pero a maiores no lexicón vinculamos estas variantes non modélicas, como facemos sempre que existen diverxencias ortográficas, mediante o hiperlema, constituído aquí polo nome científico oficial. Por exemplo:

Forma	Etiqueta	Lema	Hiperlema
Mycobacterium tuberculosis	Snaa	*Mycobacterium tuberculosis*	*Mycobacterium tuberculosis*
mycobacterium tuberculosis	Snaa	*mycobacterium tuberculosis*	*Mycobacterium tuberculosis*
M. tuberculosis	Snaa	*M. tuberculosis*	*Mycobacterium tuberculosis*

Por último, para obter un listado inicial de candidatos destinados a incrementa-lo lexicón combinamos dúas estratexias:

1. Aplicámo-la ferramenta GNRD [13] (https://gnrd.globalnames.org/) sobre quince documentos do corpus que sabiamos de antemán que contiñan denominacións científicas para detectalas e extraelas. Esta ferramenta soporta só inglés e alemán, polo que debe ser mellorada para podela aplicar con fiabilidade sobre linguas románicas. Porén, a pesar de extraer como candidatos segmentos de lingua común (*As cidades con*, *As medidas para*...), de non detectar aqueles casos que conteñen unha abreviación (*E. ensis*, *C. edule*, *Ulva sp.*, *Gracilaria spp.*, etc.) ou nos que o subtipo comeza con maiúscula (*Muellerius Capilaria*, *Protostrongilus Rufescens*...), resultou de suma utilidade, xa que entre os termos extraídos e a consulta no corpus de moitos deses candidatos singulares (formados en xeral só polo xénero) elaboramos unha lista inicial de 1014, os cales introducimos despois no lexicón. Polo tanto, na próxima versión do CORGA serán xa identificados e etiquetados como Snaa. Esta estratexia móstrase eficaz, aínda que en contrapartida require supervisión manual.
2. Inventariámo-las denominacións existentes en varias publicacións especializadas (622 aves [14], 454 árbores [15], 1023 peixes [16]) e examinamos varios repositorios, entre eles o do proxecto terminolóxico *Ictioterm* (http://www.ictioterm.es/, 358 denominacións) e o de *UniProt* (https://www.uniprot.org/). Deste último extraemos unha selección revisada constituída por 14.014 entradas (3115 bacterias, 2659 virus, 22 arqueas e 8018 eucariotas). Este método proporciona un elevado número de candidatos, pero non estamos seguros de que se vaian localizar no corpus, e pode aumentar innecesariamente o tamaño do lexicón.

Conclusión

O recoñecemento e etiquetaxe de nomes científicos en corpus anotados foi obxecto de escasa atención. Esta contribución abordou a anotación morfosintáctica automática da nomenclatura científica no CORGA, un aspecto extensible a calquera outro corpus anotado no nivel morfolóxico de calquera lingua. As claves do noso achegamento son a análise da denominación científica como unidade multipalabra, a creación dun subtipo específico dentro da clase de palabra substantivo para categorizala, a vinculación de variantes ortográficas a través do hiperlema e a ampliación do lexicón mediante a aplicación sobre o propio CORGA dunha ferramenta especializada na detección e extracción destas unidades, o cal nos permite axusta-lo lexicón ás características do corpus.