Estudos de Lingüística Galega, 14 (2022). ISSN-e: 1989-578X
https://doi.org/10.15304/elg.14.8452

Unha mellora do CORGA extrapolable a outros corpus e linguas: a etiquetaxe da nomenclatura científica binomial

An improvement for CORGA with applications to other corpora and languages: the tagging of scientific binomial nomenclature

Eva María Domínguez Noya[1,a] , Vítor Míguez[2,b]
[1] Instituto da Lingua Galega, Universidade de Santiago de Compostela / Centro Ramón Piñeiro para a Investigación en Humanidades, España
[2] Universidad del País Vasco / Euskal Herriko Unibertsitatea, España

[a]edomin@cirp.gal
[b]vitor.miguez@ehu.eus

Recibido: 05/03/2022; Aceptado: 05/10/2022

Resumo

O tratamento das unidades multipalabra é unha tarefa inconclusa no procesamento da linguaxe natural. Neste contexto, illámo-las denominacións de nomenclatura científica binomial, cuxas principais características —expresións multipalabra latinas ou latinizadas e aceptación internacional — as afastan do acervo léxico do galego e converten o seu tratamento en extrapolable a outras linguas. Tras revisa-la súa caracterización no CORGA e noutros corpus peninsulares, propoñemos analizalas como un subtipo específico de substantivos, *nomenclatura científica*, sen concretar valores de xénero nin número. Describimos logo as actuacións desenvolvidas no *kérnel* ou núcleo e mais no corpus de adestramento para integra-la nova etiqueta no sistema XIADA e, a continuación, avaliamos dúas estratexias para a detección de candidatos: unha ferramenta específica para a súa extracción e inventarios dispoñibles en Internet. Por último, á luz dos datos que proporciona o CORGA, constatamos unha presenza notable de termos científicos binomiais e demostrámo-la importancia da nova etiqueta para a súa identificación e distribución.

Palabras chave: corpus lingüístico; anotación morfosintáctica; nomenclatura científica binomial; galego.

Abstract

The treatment of multiword units is an unfinished task in natural language processing. In this context, we isolate binomial scientific nomenclature terms, whose main traits – Latin or Latinized multiword expressions and international recognition – distinguish them from the Galician 'popular' lexicon and make their treatment applicable to other languages. After reviewing their characterization in CORGA and other Peninsular corpora, we propose an analysis of scientific names as a particular subtype of nouns, namely, *scientific nomenclature*, without specifying values for gender and number. We then describe the interventions conducted on the kernel and the training corpus to include the new tag into the XIADA system and, subsequently, we asses two strategies for the detection of candidates: a specific tool for extracting scientific names and online inventories. Finally, in light of the data provided by CORGA, we verify a significant presence of binomial scientific terms and show the relevance of the new tag for their identification and distribution.

Keywords: linguistic corpus; morphosyntactic tagging; binomial scientific nomenclature; Galician.

SUMARIO

CONTENTS

1. INTRODUCIÓN

A nomenclatura científica binomial é a resposta, universalmente adoptada, ó reto taxonómico que supón a inxente diversidade de organismos vivos que habitan a Terra. As súas

Estudos de Lingüística Galega, 14 (2022). ISSN-e: 1989-578X
https://doi.org/10.15304/elg.14.8452

orixes están indisolublemente ligadas a Carlos Linneo, considerado o pai da nomenclatura binomial moderna, sen que isto impida recoñece-lo papel de figuras precursoras como a de Caspar Bauhin e, en xeral, o carácter gradual da evolución da nomenclatura científica (Pavlinov 2021). Neste sentido, hai que entender que os medios de descrición da biodiversidade ofrecidos polos sistemas de nomenclatura mostran unha evolución paralela á das ideas sobre a estrutura da biodiversidade e sobre os métodos para describila. Con todo, non nos deteremos aquí nas diferenzas entre escolas nomenclaturistas, pois non son relevantes para o presente traballo.

O principal obxectivo da nomenclatura científica é nomear de maneira unívoca cada especie de organismo coñecida e, ó tempo, agrupa-las especies pertencentes a un mesmo xénero. Deste xeito, emprégase un sistema semellante ó da onomástica (nome e apelido) no que cada denominación é o resultado da combinación de dous termos: o nome do xénero (con maiúscula inicial) e o nome da especie (en minúscula). En ocasións podemos atopar un terceiro elemento que actúa como especificador dos dous anteriores e, nestes casos, a denominación taxonómica designa unha subespecie. A nomenclatura científica facilita así a identificación exacta de calquera organismo, vivo ou extinto, e, ademais, resolve os problemas que puidesen resultar da diversidade de denominacións vernáculas. Esta variabilidade maniféstase, desde un punto de vista interlingüístico, en expresións romances como *polbo, pulpo, pop* ou *polvo* e, desde unha perspectiva intralingüística, en termos galegos como *cabala, correolo, escombro, rincha, verdel* ou *xarda*. A nomenclatura científica dá resposta a estes dous casos de multiplicidade denotativa mediante *Octopus vulgaris* e *Scomber scombrus*, respectivamente, ofrecendo así unha denominación estandarizada e universal.

No ámbito da lingüística computacional, a nomenclatura científica é estudada en relación co seu recoñecemento e extracción en textos de linguaxe técnica e, particularmente, no ámbito do recoñecemento de entidades mencionadas (NER, *Named Entity Recognition*, nas súas siglas en inglés), que ten como fin atopar e clasifica-las entidades mencionadas nun texto ou corpus. Os achegamentos NER á nomenclatura científica serven fins diversos. Por exemplo, Pafilis et al. (2013), ante o fluxo masivo de bibliografía biomédica, desenvolven ferramentas encargadas de minar textos científicos á procura de denominacións binomiais que permitan recoñecer entidades biolóxicas mencionadas, asistindo así no labor de selecciona-las contribucións bibliográficas relevantes para un determinado tema. Os nomes científicos tamén poden xogar un papel importante na busca de correspondencias entre nomes comúns de especies biolóxicas en diferentes linguas, como mostran Seideh, Fehri & Haddar (2017), que deseñan un método para establecer emparellamentos entre expresións francesas e árabes empregando os termos científicos. Unha última mostra do interese da nomenclatura no ámbito NER témola no traballo de Nguyen, Gabud & Ananiadou (2019), onde elaboran un corpus de entidades mencionadas na bibliografía sobre biodiversidade co obxectivo de rexistra-la ocorrencia de especies biolóxicas, especialmente aquelas ameazadas ou en perigo de extinción.

Fronte ó interese suscitado no eido NER, o recoñecemento e etiquetaxe da nomenclatura binomial en corpus lingüísticos foi obxecto de escasa atención. É máis, malia que a identificación e o tratamento das unidades multipalabra hai anos que é foco de interese (Calzolari et al. 2002, Villavicencio et al. 2007 ou Caseli et al. 2009), pódese considerar que esta é aínda unha tarefa pendente do procesamento da linguaxe natural, como apuntan Darriba, Doval & Kuriyozov (2021). Estes autores, que tentan demostra-la utilidade de técnicas de aprendizaxe profunda para a detección e identificación de expresións multipalabra empregando o CORGA, que na altura non incluía aínda a mellora na anotación da nomenclatura científica, teñen en conta no seu estudo dous grandes bloques: entidades mencionadas (isto é, *Universidade de Vigo, Fondo Monetario Internacional, Xunta de Galicia,*

Unión Europea) e locucións adverbiais, adxectivas, conxuntivas, prepositivas e substantivas (isto é, *sen dúbida, azul celeste, e mais, cara a, visto e prace*).

O CORGA é un corpus documental aberto representativo da lingua galega actual, codificado en XML e enriquecido con anotación morfosintáctica mediante a ferramenta XIADA, un etiquetador de tipo estatístico, cuxo obxectivo é fornecer datos para o estudo do galego dende múltiples perspectivas: léxica, morfolóxica, sintáctica, fraseolóxica, terminolóxica, comunicativa etc. Na actualidade pode afirmarse que o CORGA acadou un grao de madurez importante, pero aínda admite melloras en moi distintos aspectos. Estas melloras poden dirixirse, seguindo a idea exposta por Bunge (1972), ó crecemento en superficie e ó crecemento en profundidade. O crecemento en superficie consiste no incremento do volume de textos codificados e anotados e, no caso do rexistro oral, tamén na súa transcrición e aliñamento. O crecemento en profundidade, co que se relaciona o presente traballo, inclúe melloras en tódolos procesos, especialmente nos relativos á etiquetaxe morfosintáctica e ás posibilidades de explotación.

Coma calquera etiquetador, o de XIADA aplica un conxunto de etiquetas, o etiquetario, ó corpus obxecto de anotación, o CORGA. En tanto que etiquetador estatístico, faino fundamentalmente a partir dunha base de datos léxica na que se asocia a cada entrada unha etiqueta e un lema, así como dun corpus de adestramento que proporciona o modelo de desambiguación. Completa estes módulos outro de regras lingüísticas que permite mellorar a taxa de acerto (Domínguez Noya 2016).

Na primeira capa de análise lingüística aplicada sobre o CORGA (Domínguez Noya 2013) non se tiveron 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. Isto entronca directamente coa convención que, no establecemento da nomenclatura científica para as especies, determina escribi-la inicial do xénero en maiúscula, o cal conduce na etiquetaxe automática das versións do CORGA anteriores á 4.0 a unha análise desagregada errónea da expresión binomial (isto é, *Octopus vulgaris*), na que a forma correspondente ó xénero se caracteriza como substantivo propio e recibe como lema o propio nome (*Octopus/Sp00/Octopus*), mentres que a parte relativa ó subtipo adoita clasificarse como substantivo ou adxectivo (para o exemplo, substantivo común masculino/feminino plural, sen lema: *vulgaris/Scap/**). Son precisamente estas caracterizacións incorrectas a xeito de unidades independentes as causantes de que nos acheguemos ás expresións binomiais da nomenclatura científica, co fin de favorece-la súa anotación conxunta, integrándoas de modo coherente coas características dun recurso que conta xa con algo máis de 40 millóns de palabras (43 162 364 palabras ortográficas / 51 451 088 elementos gramaticais).

Nesta contribución, tras revisa-lo seu tratamento noutros corpus peninsulares, ilustrámo-lo proceso mediante o cal engadímo-la nomenclatura binomial ó repertorio de expresións multipalabra do CORGA coa esperanza de que poida resultar útil tanto para a súa futura inclusión en estudos de procesamento da linguaxe natural que empreguen este corpus como para a súa aplicación noutros corpus doutras linguas.

2. A ANOTACIÓN DOS TERMOS CIENTÍFICOS NOUTROS CORPUS

A consulta doutros corpus anotados, de referencia ou técnicos, do galego (*Tesouro Informatizado da Lingua Galega* [TILG] e *Corpus Técnico Anotado do Galego* [CTAG]), español (*Corpus del Español del Siglo XXI* [CORPES]), portugués (*Corpus de Referência do Português*

Estudos de Lingüística Galega, 14 (2022). ISSN-e: 1989-578X
https://doi.org/10.15304/elg.14.8452

Contemporâneo [CRPC]) e catalán (*Corpus Tècnic* [CT] e *Corpus Textual Informatitzat de la Llengua Catalana* [CTILC]) presenta resultados heteroxéneos, como amosa a selección recollida na táboa 1.

Táboa 1. Análise de termos científicos binomiais no TILG, CTAG, CORPES, CRPC, CT e CTILC

TERMO BINOMIAL	TILG	CTAG	CORPES	CRPC	CT	CTILC
QUERCUS ROBUR; *QUERCUS ILEX** (CARBALLO; ACIÑEIRA)	2 unidades: Palabra estranxeira + Palabra estranxeira	1 unidade: NCMS0 (subs. común masc. sg.)	2 unidades: estranxeirismo + estranxeirismo	2 unidades: PNM + PNM (parte do nome)	*1 unidade: N5-66 (nome común, xénero e número pendentes)	1 unidade: NC (non codificado)
CASTANEA SATIVA (CASTIÑEIRO)	2 unidades: Palabra estranxeira + Palabra estranxeira	1 unidade: NCFS0 (subs. común fem. sg.)	2 unidades: estranxeirismo + estranxeirismo	2 unidades: PNM + V#ii-3s (3.ª sg. do copretérito de indicativo)	1 unidade: N5-66	1 unidade: NC
SUS SCROFA (XABARIL)	2 unidades: Palabra estranxeira + Palabra estranxeira	1 unidade: NCFS0 (subs. común fem. sg.)	2 unidades: posesivo, masc. pl. (lema suyo) + ?? (descoñecido)	2 unidades: PNM + PNM ou PNM + V#pi-3s (3.ª sg do presente de indicativo)	1 unidade: N5-66	1 unidade: NC
*PINUS PINASTERPINUS SILVESTRIS** (PIÑEIRO)	2 unidades: Palabra estranxeira + Palabra estranxeira	1 unidade: NCMS0 (subs. común masc. sg.)	2 unidades: Subs. propio + estranxeirismo	2 unidades: PNM + PNM	*1 unidade: N5-66	1 unidade: NC
BETULA ALBA; *BETULA PUBESCENS** (BIDUEIRO)	*2 unidades: Palabra estranxeira + Palabra estranxeira	1 unidade: NCFS0 (subs. común fem. sg.)	2 unidades: Subs. propio + estranxeirismo	2 unidades: PNM + V#mpi-3s (3.ª sg. do antepretérito ind.)	----------	1 unidade: NC

Dos corpus examinados, entre os que, ademais dos referenciados na táboa 1, cómpre engadirmos o *British National Corpus* (BNC), o *Corpus Brasileiro* (CB) e mailo *Corpus del español* (CdE) e o *Corpus do português* (CdP) de Mark Davies, ningún reserva unha etiqueta específica para a nomenclatura científica e só o CTAG e mailos dous do catalán, CT e CTILC, 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 analiza os membros por separado e a súa categorización tende á que amosa o CORPES: estranxeirismo ou substantivo propio para o xénero, e estranxeirismo ou descoñecido para o subtipo. Comentario á parte merece, sen dúbida, o CRPC, que amosa as análises máis sorprendentes, inda que xustificables por se-la etiquetaxe automática, ó caracteriza-lo termo para o subtipo *sativa, scrofa* e *alba* como forma verbal de 3.ª persoa singular indicativa en copretérito, presente e antepretérito, respectivamente.

3. PROPOSTA DE ANÁLISE

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. Propoñemos, polo tanto, a 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.

Á luz do tratamento unitario que o CTAG, o CT e o CTILC aplican sobre as denominacións científicas binomiais, así como da súa caracterización, cremos conveniente tratalas tamén no CORGA como substantivos (S). Con todo, optamos por crear un novo subtipo que denominamos *nomenclatura científica* (n) e non integralos entre os substantivos comúns. Tal decisión alicérzase en dous motivos. O primeiro deles pode sintetizarse en que estas expresións pertencen a unha lingua distinta do galego: son expresións latinas ou latinizadas que non forman parte, por tanto, do seu acervo léxico; como tales, ademais, fóra algún caso de uso xa común, non se recollen nas obras lexicográficas que inventarían o léxico dunha lingua. Por último, a mesma relación de termos científicos que denominan especies ou subespecies é de uso no galego, no catalán, no español ou no portugués, por só citar algunhas das linguas posibles, sen necesidade de adaptación. En consecuencia, non estamos ante un trazo intrínseco a unha lingua específica, cuxo comportamento gramatical difire en función da lingua de que se trate, como ocorre, por exemplo, coa flexión do número para os substantivos agudos polisílabos rematados en *-n*. É precisamente esta universalidade a que permite extrapola-la presente proposta aplicada no CORGA para o galego a outros corpus e linguas. O segundo dos motivos fundaméntase na practicidade: a creación dun valor exclusivo para a nomenclatura científica individualiza estes elementos mediante unha etiqueta específica e facilita a súa recuperación e extracción na plataforma de consulta.

En relación, finalmente, cos trazos gramaticais de xénero e número pertinentes para a clase de palabra substantivo, as denominacións binomiais aparecen frecuentemente en contextos lingüísticos pobres, onde os atributos gramaticais propios dos substantivos non se manifestan: como primeiro elemento da entrada dun apéndice tipo dicionario (“Quercus suber L. = Sobreira.”); como aclaración dun nome común (“dereita: un delicioso rolo de kelp de AMTI (Saccharina latissima) preparado por Chris Aerni [...]”); en aposición a un substantivo común (“Nas zonas protexidas, a especie Fucus vesiculosus viuse fortemente danada [...]”, “Na páxina anterior, unha colleiteira nas costas galegas da alga Gelidium corneum utilizada pola industria dos ficocoloides.”); en enumeracións (“Proponse a separación da vexetación en tres tramos altitudinais, caracterizado cada un deles por unha especie dominante sobre a paisaxe vexetal: o piñeiro (Pinus pinaster), o castiñeiro (Castanea vulgaris), o carballo (Quercus pedunculata) e o xenebro (Juniperus nana), segundo unha orde de altitudes crecentes.”); etc.

Con todo, non é raro atopar no CORGA expresións binomiais en contextos lingüisticamente máis ricos nos que se dá algún tipo de determinación e/ou modificación: “Na Galiza tamén podemos atopa-la Dactylorhiza sambucina.”, “É moi semellante ao Boletus edulis, do que se diferenza polo tamaño, é algo máis pequeno.”, “Especial interese, cara á quimioprofilaxe e ó tratamento, teñen os P. Falciparum resistentes á cloroquina e, polo tanto, teno tamén o coñecemento da súa distribución xeográfica.”, “En principio, a localización das Muscari neglectum atopadas, non facía prever un dano próximo a esta pequena poboación.”.

Aínda así, seguindo o CT e o CTILC, decidimos non concreta-los valores de xénero e número para o subtipo *nomenclatura científica*. O feito de apareceren maioritariamente illados en unidades parentéticas, de seren termos non galegos, de non contarmos con coñecementos suficientes de latín, de non dispormos de glosarios coa información gramatical xenérica oportuna e de requirir un labor de investigación que, en termos de tempo e recursos, iría en detrimento do proxecto conducen ó establecemento de hipervalores en ámbolos dous atributos[1]. Así, na terceira posición dos caracteres que constitúen a abreviación da etiqueta,

Estudos de Lingüística Galega, 14 (2022). ISSN-e: 1989-578X
https://doi.org/10.15304/elg.14.8452

'Snaa', o 'a' equivale ó valor inespecificado 'masculino/feminino', mentres que o 'a' da cuarta posición corresponde ó valor 'singular/plural'. A seguir descríbense os procesos requiridos para aplicar na anotación morfosintáctica a nova etiqueta.

4. IMPLEMENTACIÓN DA NOVA ETIQUETA NO SISTEMA XIADA

O etiquetador de XIADA, coma calquera outro, aplica o etiquetario definido ó corpus obxecto de anotación, o CORGA. En tanto que etiquetador estatístico, faino principalmente a partir dunha base de datos léxica na que se asocia a cada entrada unha etiqueta e un lema, así como dun corpus de adestramento a partir do cal infire a información gramatical que lle proporciona o modelo de desambiguación. Estes módulos compleméntanse con outros que melloran a taxa de acerto ou inciden sobre a reconstrución do elemento gramatical cando este está inmerso, como palabra ortográfica, nunha amálgama.

Para implementa-la etiqueta 'Snaa' en XIADA non basta con introduci-lo valor nomenclatura científica no subtipo dos substantivos[2] nin abonda con suma-la etiqueta e un exemplo de uso á listaxe de etiquetas que poden aparecer asignadas no CORGA[3]. Estas son tarefas exclusivamente de documentación que non repercuten na etiquetaxe do corpus: só informan das clases de palabras recoñecidas e mais dos trazos morfosintácticos que as caracterizan, xunto con exemplos de uso.

En realidade, para poder empregala, o etiquetador só esixe que a etiqueta estea recollida no que en lingüística de corpus se denomina *Training Corpus Zero*, ó cal imos referirnos de aquí en diante como *kérnel* ou núcleo de XIADA. Se se cumpre esta condición, a nova etiqueta 'Snaa' formará xa parte do xogo de candidatos que a ferramenta de anotación manexará para caracteriza-las palabras do texto que se pretenda procesar. Agora ben, para que a anotación automática poida realizarse con certas garantías de éxito, un etiquetador de tipo probabilístico, como é o de XIADA, debe contar tamén con información sobre cando esa etiqueta posúe máis probabilidades de se-la que corresponde aplicar e debe, por último, dispoñer dunha relación de unidades susceptibles de recibir tal caracterización morfosintáctica. Ou sexa, hai que intervir tamén sobre o corpus de adestramento e o lexicón. Nos subapartados que seguen, desenvolvemos con máis detalle as actuacións realizadas, respectivamente, sobre estes tres módulos: *kérnel*, corpus de adestramento e lexicón.

4.1. Introdución no núcleo de XIADA

Como xa sinalamos, o único requisito para poder executar o etiquetador de XIADA é dispoñer do *kérnel* ou núcleo no que debe constar cada etiqueta presente no etiquetario analizada como mínimo unha vez. O núcleo de XIADA está constituído por un arquivo de só texto no que se recolle a análise de 332 enunciados (5513 palabras ortográficas / 6629 elementos gramaticais).

O método escollido para reflecti-las análises é o seguinte: as secuencias descompóñense nos seus elementos constituíntes, de xeito que as contraccións e conglomerados de formas verbais con segunda forma do artigo ou pronomes enclíticos se desagregan nos distintos elementos gramaticais dos que constan; por exemplo, *dos* sepárase en *de* e *os*, de modo que cada un deles ocupa unha liña; pola contra, as palabras que conforman unha unidade multipalabra, tipo *garda civil, Centro Galego de Arte Contemporánea, vintesete mil trescentos* ou *con respecto a*, recóllense na mesma liña. A continuación, en cada unha desas liñas conformadas polo elemento gramatical, introdúcese a información relativa á etiqueta que corresponda segundo o contexto no que se localice e remítese ó seu lema. Ou sexa, cada liña

recolle a información relativa a unha forma-etiqueta-lema, separando os valores de cada campo cun tabulador. Á súa vez, unha liña en branco separa as análises das distintas secuencias.

Os enunciados, sempre que amosaban a análise correspondente a unha ou a varias das etiquetas cuxo uso se pretendía ilustrar, tiráronse do propio CORGA. Cando isto non era posible por non ser usos claros, pola extensión do fragmento no que se localizaba ou pola dificultade de documentala, optouse por exemplos inventados. Este mesmo proceder seguiuse para exemplifica-la nova etiqueta: procuramos no corpus os casos de *Octopus vulgaris*, seleccionamos de entre os resultados a secuencia que recollemos deseguido, a cal presenta catro casos de nomenclatura científica binomial (*Paracentrotus lividus, Pollicipes pollicipes, Necora puber, Octopus vulgaris*), e procedemos a analizala dun xeito manual no formato descrito anteriormente: "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." (CORGA: Xaquín Penas Patiño, *A arca de Noé. Elementos de Oceanografía e Bioloxía mariña*, Baía Edicións, 2006).

4.2. Revisión do corpus de adestramento

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.

Como sinala Graña (2000) e, seguindo a súa senda, Domínguez Noya (2013, 2016), na anotación morfosintáctica cun etiquetador de tipo probabilístico hai que ter en conta o tamaño e calidade do corpus de adestramento (a maior tamaño e mellor calidade, mellor aprendizaxe do modelo), o etiquetario e a base de datos léxica. Este feito corrobórao tamén Manning (2011), quen sinala que se acada a precisión máxima en etiquetadores deste tipo cando hai congruencia na temática, época e estilo entre os datos presentes no corpus de adestramento e aqueles que se pretenden anotar. Cómpre, polo tanto, que haxa consistencia entre o corpus de adestramento e o corpus anotado dende o punto de vista do rexistro e o xénero textual, necesidade recoñecida por todos estes autores, mais tamén é fundamental que estea etiquetado coidadosamente, de xeito que a mesma unidade en contextos similares reciba a mesma caracterización.

Así pois, malia que nada garante que no subcorpus seleccionado para servir de adestramento se vaian localizar tódalas etiquetas contidas no etiquetario, cómpre que haxa consistencia entre os recursos, polo que é imprescindible testar, e reanotar se for preciso, a posible presenza dalgún termo científico binomial ou trinomial.

Para levar a cabo esta tarefa desbotouse a revisión manual do corpus debido, por unha banda, ó seu tamaño (617 494 palabras ortográficas, 742 045 elementos gramaticais), mais tamén, por outra banda, debido á innecesariedade de tal labor: o corpus de adestramento está dispoñible para a súa consulta en liña[4] e a aplicación permite recuperar e inventariar, grazas á combinación da información relativa a unha parte da etiqueta (Sc* [substantivo común] e Sp* [substantivo propio]), co metacáracter que substitúe un, ningún ou varios caracteres (* * no campo do elemento gramatical), tódolos substantivos multipalabra formados por, como mínimo, dous elementos gramaticais, tanto comúns coma propios, co que obtivemos tódolos candidatos. Foi preciso, iso si, unha vez obtidos os datos, revisar manualmente ámbolos dous inventarios e acudir ás concordancias cando o resultado non era claro, como ocorre con *Pittacum Barrica* que, malia te-la aparencia de termo científico binomial, é unha marca de viño do Bierzo.

Estudos de Lingüística Galega, 14 (2022). ISSN-e: 1989-578X
https://doi.org/10.15304/elg.14.8452

Do total de resultados recuperados (5434 e 121 *types*[5] analizados como substantivos propios e comúns, respectivamente) localizamos no subcorpus de adestramento entre os comúns *ginkgo biloba* (3), *piñeiro pinaster* (2) e *estafilococusaureus* (1), mentres que entre os propios detectamos *Caenorhabditis elegans* e *Tursiops Truncatus* con só unha ocorrencia cada un. Destes, entendemos que só os dous últimos debían ser etiquetados como nomenclatura científica, mentres que os demais debían conserva-la etiqueta de substantivo común. O primeiro, *ginkgo biloba*, por se-la denominación común dunha especie foránea e carecer da maiúscula inicial; o segundo, tamén pola ausencia da maiúscula inicial no xénero e a mestura do termo común co científico (*piñeiro pinaster* en vez de *Pinus pinaster*), e, por último, o "e" inicial minúsculo en "estafilococus", e maila ausencia neste do "ph" e "cc" presentes na denominación científica convencional: *Staphylococcus aureus*.

Táboa 2. Mostra das irregularidades presentes en denominacións científicas no CORGA

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

Como se constata, a relación anterior non está exenta de controversia e exemplifica as dificultades que supón traballar con textos reais (Táboa 2) que a miúdo conteñen erros tipográficos (a-c), caracteres alleos ó latín (d-e) ou usos incorrectos das maiúsculas (f-g).

Táboa 3. Mostra de variación no lemario de XIADA para a mesma denominación científica

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*

En todos estes casos establecemos como lema a propia expresión binominal documentada no CORGA, pero, a maiores, vencellamos na base de datos léxica estas variantes non modélicas, como sempre que existen diverxencias ortográficas entre as distintas entradas do dicionario, mediante o hiperlema, constituído aquí polo nome científico oficial (Táboa 3).

4.3. Introdución de candidatos no lexicón

En último termo, para obter un listado inicial de candidatos destinados a incrementa-lo lexicón combinamos dúas estratexias.

Por un lado, aplicámo-la ferramenta *Global Names Recognition and Discovery* GNRD[6] (Pyle 2016), sobre dezaseis documentos do corpus[7] 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 maila consulta no corpus de moitos deses candidatos singulares, formados en xeral só polo xénero, elaboramos unha lista inicial de 1014, os cales se introduciron logo no lexicón. Como veremos na seguinte epígrafe, esta estratexia móstrase eficaz, aínda que en contrapartida require supervisión manual.

Por outra parte, inventariámo-las denominacións existentes en varias publicacións especializadas (622 aves [Rouco et al. 2019], 454 árbores [Rivers 2019], 1023 peixes [Resolución 2019]) e examinamos varios repositorios, entre eles o do proxecto terminolóxico *Ictioterm*[8] (358 denominacións) e o de *Uni-Prot*[9]. Deste último extraemos unha selección revisada constituída por 14 014 entradas (3115 bacterias, 2659 virus, 22 arqueas e 8018 eucariotas). En síntese, reunimos 29 994 termos de nomenclatura científica binomial ou trinomial, a razón de un por liña, que se lle proporcionaron ó etiquetador externamente, sen chegar a introducilos no lexicón. Este método permite obter, con relativa facilidade, un elevado número de candidatos, mais, en contrapartida, nada garante que estes se vaian localizar no corpus, co cal pode aumentar innecesariamente o tamaño do lexicón.

5. EFICACIA DA NOVA ETIQUETA

A implementación da etiqueta 'Snaa' no sistema XIADA actúa nun dobre sentido: por un lado, facilita a recuperación global e/ou individualizada de casos de nomenclatura científica binomial ou trinomial presentes no CORGA e abre as portas a estudos relacionados coa súa presenza nun xénero textual específico, coa súa frecuencia de uso, posibles causas etc.; por outro lado, supón unha mellora substancial na etiquetaxe do propio corpus por canto reduce os casos mal identificados e mal anotados que proporcionaban as versións previas, como sucedía coa análise de *Octopus vulgaris* e outros moitos, descrita no apartado introdutorio.

Entre a información estática que fornece o CORGA é de interese para o tema que nos ocupa a listaxe da frecuencia das etiquetas[10], posto que permite obter con suma facilidade o dato da frecuencia no total do etiquetario para 'Snaa'. Así, das 454 etiquetas diferentes ás que dá lugar o *tagset* de XIADA, no corpus etiquetado automaticamente concorren 419, e destas, a de 'Snaa' ocupa a posición 222, cun total de 2354 ocorrencias, das cales son formas distintas 1234. Recollemos como mostra deseguido os vinte termos científicos binomiais máis frecuentes, xunto co número de ocorrencias global (frecuencia absoluta) e mailo número de documentos diferentes no que se localizan. Son os seguintes: *E. coli* (60/9), *Homo sapiens* (34/11), *Saccharina latissima* (33/8), *Romulea columnae* (19/2), *Homo erectus* (18/5), *Barlia robertiana* (15/2), *Eucalyptus glóbulus* (13/1), *Pinus pinaster* (13/7), *Boletus edulis* (12/5), *Escherichia coli* (12/6), *P. olseni* (12/1), *Centaurea borjae* (11/2), *Eucalyptus globulus* (11/6), *Homo habilis* (11/4), *Quercus robur* (11/8), *Castanea sativa* (10/5), *Digitalis purpurea* (10/5), *D. magna* (10/1), *Ensis siliqua* (10/4) e *Ostrea edulis* (10/7).

A recente actualización do CORGA permitiu examina-la eficacia dos dous métodos empregados para a detección e identificación dos termos científicos binomiais. Primeiramente, en novembro de 2021, realizouse a etiquetaxe do corpus, no que atinxe á nomenclatura científica, coa información proporcionada só polo lexicón (1014 termos de nomenclatura científica), o cal posibilitou o recoñecemento de 5446 ocorrencias, das cales eran formas distintas 2459. Unhas cifras, en principio, inesperadamente altas. Non obstante, a observación dos resultados amosou formas simples ás que o módulo de adiviñación lles

asignaba a etiqueta 'Snaa', entendemos que pola semellanza na terminación, entre as cales presentaban maior frecuencia absoluta as seguintes: *princepessa, manager, bossa, europeus, hutus, confeti, magnum, henna* ou *valium*. É dicir, o *guesser*, que nunca actúa sobre unidades multipalabra, aplicaba a etiqueta erroneamente, polo que decidimos excluíla de entre as que manexa. Volveuse anotar todo e obtivéronse 1812 ocorrencias, das cales 881 eran formas distintas, mais agora si, todos os resultados correspondían a expresións binomiais científicas. Un mes máis tarde reanótase de novo todo o corpus, mais neste caso tendo en conta non só a información do lexicón, senón tamén a do arquivo externo que contén case 30 000 candidatos colleitados en repositorios de Internet. Os datos, coincidentes xa cos que achega a versión 4.0 dispoñible en liña, ofrecen agora un resultado de 2468 ocorrencias identificadas, das cales 1346 son formas únicas; ou sexa, o recoñecemento do 73% prodúcese grazas ó lexicón computacional.

En suma, este parece se-lo camiño a seguir: empregala ferramenta GNRD sobre o propio corpus para a detección de candidatos, executa-lo etiquetador sobre o resultado e introducir no lexicón os non recoñecidos, tras verifica-la súa pertinencia.

6. CONCLUSIÓNS

No ámbito do procesamento da linguaxe natural o tratamento das unidades multipalabra é aínda un labor inconcluso. Neste contexto, encaramos un problema sistemático de anotación morfosintáctica no CORGA, o cal se produce tamén noutros corpus, de referencia ou técnicos, etiquetados automaticamente: a etiquetaxe desagregada e errónea da denominación científica para as especies e subespecies. Revisámo-la súa anotación no TILG, no CTAG, no CORPES, no CRPC, no CT e no CTILC. Propuxemos (i) considerar que as denominacións científicas latinizadas, binomiais ou trinomiais, conforman unha unidade multipalabra; (ii) determinámo-la súa integración na clase de palabra substantivo, (iii) dentro do subtipo *nomenclatura científica* para facilita-la súa recuperación na plataforma de consultas, ó tempo que (iv) establecemos como lema a propia forma e, en último termo, (v) vencellámo-las variantes ortográficas entre os lemas dunha mesma especie mediante a remisión ó mesmo hiperlema. Describimos así mesmo a inclusión da nova etiqueta ('Snaa') no sistema XIADA e afrontámo-la identificación dos candidatos dende unha dupla perspectiva: colleitalos en inventarios accesibles na web e extraelos do propio corpus cunha ferramenta específica de detección (GNRD). Os resultados iniciais amosan unha presenza notable deste tipo de termos no corpus e sinalan a última aproximación como a máis axeitada por canto, sen sobrecarga-lo lexicón, conduce ó recoñecemento das expresións binomiais presentes no CORGA. Finalmente, cómpre subliña-la importancia da nova etiqueta na recuperación selectiva dos datos e a adaptabilidade da presente proposta para a súa aplicación noutros corpus.

Recursos electrónicos

BNC: *British National Corpus (XML edition)*https://cqpweb.lancs.ac.uk> [Consultado: 9/2/2022]

CB: *Corpus Brasileiro* <http://corpusbrasileiro.pucsp.br/cb/Inicial.html> [Consultado: 9/2/2022]

CdE: *Corpus del español (Género/Histórico)* <https://www.corpusdelespanol.org/hist-gen/> [Consultado: 9/2/2022]

CdP: *Corpus do português (Género/Histórico)* <https://www.corpusdoportugues.org/hist-gen/> [Consultado: 9/2/2022]

CORGA: *Corpus de Referencia do Galego Actual (CORGA)* <http://corpus.cirp.gal/corga> [Consultado: 1-17/2/2022]

CORPES: *Corpus del Español del Siglo XXI*. <https://www.rae.es/> [Consultado: 9/2/2022]

CRPC: *Corpus de Referencia do Português Contemporâneo*. <http://alfclul.clul.ul.pt/CQPweb/crpcfg16/> [Consultado: 9/2/2022]

CT: *Corpus Tècnic*. <https://www.upf.edu/es/web/iula/corpus> [Consultado: 9/2/2022]

CTAG: *Corpus Técnico Anotado do Galego*. <http://sli.uvigo.es/CTAG/> [Consultado: 9/2/2022]

CTILC: *Corpus textual informatitzat de la llengua catalana*. <https://ctilc.iec.cat/> [Consultado: 9/2/2022]

TILG: *Tesouro informatizado da lingua galega*. <http://ilg.usc.es/TILG/> [Consultado: 9/2/2022]

XIADA: *Etiquetador/Lematizador do Galego Actual (XIADA) [2.8]* <http://corpus.cirp.gal/XIADA>

Referencias bibliográficas

Bunge, Mario. 1972. *La investigación científica*. Barcelona: Ariel.

Calzolari, Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod & Antonio Zampolli. 2002. Towards Best Practice for Multiword Expressions in Computational Lexicons. En Manuel González Rodríguez & Carmen Paz Suarez Araujo (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02).* 1934-1940. Las Palmas: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2002/pdf/259.pdf

Caseli, Helena, Aline Villavicencio, André Machado & Maria José Finatto. 2009. Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains. En Dimitra Anastasiou, Chikara Hashimoto, Preslav Nakov & Su Nam Kim (eds.), *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)*. 1-8. Singapore: Association for Computational Linguistics. https://aclanthology.org/W09-2901.pdf

Darriba, Víctor, Yerai Doval & Elmurod Kuriyozov. 2021. Procesamiento de expresiones multipalabra en gallego mediante Aprendizaje Profundo. *Procesamiento del Lenguaje Natural, 67*, 45-57. https://doi.org/10.26342/2021-67-4

Domínguez Noya, Eva María. 2013. *Etiquetaxe e desambiguación automáticas en galego: o sistema XIADA*. Santiago de Compostela: Universidade de Santiago de Compostela. [Tese de doutoramento inédita]. http://hdl.handle.net/10347/9587

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

Ernout, Alfred & Antoine Meillet. 2001. *Dictionnaire étymologique de la langue latine. Histoire des mots*. Paris: Klincksieck. [Obra publicada orixinalmente en 1932].

Graña Gil, Jorge. 2000. *Técnicas de análisis sintáctico robusto para la etiquetación del lenguaje natural*. A Coruña: Universidade da Coruña. [Tese de doutoramento inédita]. http://hdl.handle.net/2183/12358

Manning, Christopher D. 2011. Part-of-speech tagging from 97 % to 100 %: is it time for some linguistics?. En Alexander F. Gelbukh (ed.), *Computational linguistics and intelligent text processing, 12th International Conference, CICLing 2011, Proceedings*. Part I: *Lecture notes in computer science 6608*. 171-189. Berlin: Springer. https://doi.org/10.1007/978-3-642-19400-9_14

Nguyen, Nhung T. H., Roselyn S. Gabud & Sophia Ananiadou. 2019. COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal* 7, e29626. https://doi.org/10.3897/BDJ.7.e29626

Pafilis, Evangelos, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis & Lars Juhl Jensen. 2013. The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoSONE* 8(6), e65390. https://doi.org/10.1371/journal.pone.0065390

Pavlinov, Igor Ya. 2021. *Taxonomic nomenclature: What's in a name – theory and history*. Boca Raton: CRC Press. https://doi.org/10.1201/9781003182535

Pyle, Richard L. 2016. Towards a Global Names Architecture: The future of indexing scientific names. *ZooKeys* 550, 261-281. https://doi.org/10.3897/zookeys.550.10009

Resolución de 24 de mayo de 2019, de la Secretaría General de Pesca, por la que se publica el listado de denominaciones comerciales de especies pesqueras y de acuicultura admitidas en España, *Boletín Oficial del Estado*, 143, de 15/06/2019. https://www.boe.es/buscar/doc.php?id=BOE-A-2019-9026

Rivers, Malin. 2019. *European Red List of trees*. Cambridge / Brussels: IUCN. https://doi.org/10.2305/IUCN.CH.2019.ERL.1.en

Rojo, Guillermo. 2017. Sobre la configuración estadística de los corpus textuales. *Lingüística* 33(1), 121134. http://doi.org/10.5935/2079-312x.20170008

Rouco, Miguel, José Luis Copete, Eduardo de Juana, Marcel Gil-Velasco, Juan Antonio Lorenzo, Marce Martín, Borja Milá, Blas Molina & David M. Santos. 2019. *Lista de las aves de España*.

Madrid: SEO/BirdLife. https://seo.org/wp-content/uploads/2019/05/ListaAvesdeEspa%C3%B1a2019.pdf

Seideh, Mohamed Aly Fall, Hela Fehri, & Kais Haddar. 2017. Recognition and extraction of Latin names of plants for matching common plant named entities. En Linda Barone, Mario Monteleone & Max Silberztein (eds.), *Automatic processing of natural-language electronic texts with NooJ. 10th International Conference, NooJ 2016, České Budějovice, Czech Republic, June 9-11, 2016, Revised Selected Papers*. 132-144. Berlin: Springer. https://doi.org/10.1007/978-3-319-55002-2_12

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart & Carlos Ramisch. 2007. Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. En Jason Eisner (ed.), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 1034-1043. Prague: Association for Computational Linguistics. https://aclanthology.org/D07-1110.pdf

Notas

[1] O xénero gramatical non pode ser atribuído tendo en conta as terminacións, pois levaría a erros como o que cometeriamos ó supoñer que os rematados en -us, tipo *Pinus pinaster ou Quercus robur*, son masculinos, cando os dicionarios e gramáticas latinas establecen para os nomes de árbores a excepcionalidade da regra que sinala que os rematados en -us da segunda declinación son en xeral masculinos (Ernout & Meillet 2001).

[2] http://corpus.cirp.gal/XIADA/etiquetario/taboa

[3] http://corpus.cirp.gal/XIADA/etiquetario/exemplos

[4] Opción *Etiquetado manualmente* para o ítem *Corpus* situado no bloque *Busca* da pantalla de captación de datos da aplicación de consulta do CORGA.

[5] Seguindo a Rojo (2017), distinguimos entre tokens (formas dun texto) e *types* (formas distintas). Ó poder reduci-las ocorrencias a formas distintas simplifícase a tarefa, pois non nos vemos obrigados a revisar, un supoñer, 15 casos de *conta corrente*, senón só $_{1}$.

[6] Dispoñible no enderezo https://gnrd.globalnames.org/.

[7] As obras referidas son os ensaios *A ciencia do pole en Galicia*. *Metodoloxía e Aplicacións*; *A cuestión ambiental en Galicia*. *Raíces dunha nova cultura 1750-1972*; *Acuicultura multitrófica integrada*. *Unha alternativa sustentable e de futuro para os cultivos mariños en Galicia*; *Antioxidantes naturais*. *Aspectos saudables*, *toxicolóxicos e aplicacións industriais*; *As plantas medicinais*; *De Darwin ao ADN*. *Ensaios sobre as implicacións sociais da Bioloxía*; *Ecoloxía forestal e ordenación do bosque*; *Enfermidades infecciosas emerxentes*; *Estudo integral das Rías Altas*. *Estudo das rías de Ribadeo*, *Foz*, *Viveiro*, *O Barqueiro e Ortigueira*. *Hidrografía*, *dinámica*, *bioxeoquímica*, *sedimentoloxía*, *ecotoxicoloxía*, *microbioloxía, patoloxía e bioloxía das zonas de interese marisqueiro*; *Festas gastronómicas de Galicia*. *Festas, receitas, calendario e puntos de interese turístico*; *Guía dos mariscos de Galicia*; *Manual marítimo-pesqueiro*: *Galicia*; *Pesca submarina en Galicia e Unha alternativa gandeira pra os montes da Galiza* e, por último, o libro de texto para o segundo curso de bacharelato de *Bioloxía* e mais o blog *Natureza Dixital. Retallos de natureza en formato dixital*.

[8] Véxase http://www.ictioterm.es.

[9] Véxase https://www.uniprot.org/.

[10] Dispoñible en http://corpus.cirp.gal/corga/frecuencias/etiquetado%20automaticamente.