
María Álvarez de la Granja / Marta Negro Romero (2015): "O processo de lematização no Tesouro do léxico patrimonial galego e português", en Fabiane Cristina Altino / Gleidy Aparecida Lima Milani / Rosa Evangelina Santana Belli Rodrigues (coords.), *Anais do III CIDS: Congresso Internacional de Dialetologia e Sociolinguística*. Londrina (Brasil): Universidade Estadual de Londrina, 848-862.

ANAIS DO III CIDS
Congresso Internacional de Dialetologia e Sociolinguística
– Variedade, atitudes linguísticas e ensino

Homenagem a Jacyra Andrade Mota e Suzana Alice Marcelino Cardoso

Centro de Letras e Ciências Humanas (CLCH)
Universidade Estadual de Londrina – 07 a 10 de outubro de 2014

ISBN 978-85-7846-344-1

O PROCESSO DE LEMATIZAÇÃO NO *TESOURO DO LÉXICO PATRIMONIAL GALEGO E PORTUGUÊS*[1]

Trabalho apresentado no Simpósio
"Lexicografia dialetal na era digital: o Tesouro do Léxico Patrimonial Galego e Português"

María **Álvarez de la Granja***
Universidade de Santiago de Compostela
Marta **Negro Romero** (PG)**
Universidade de Santiago de Compostela

Resumo: O *Tesouro do Léxico Patrimonial Galego e Português* é uma base de dados léxica que permite o acesso à informação contida em trabalhos de léxico dialetal do galego, do português de Portugal e do português do Brasil. Os dados das fontes originais oferecem-se completos e organizados a partir de variantes, lemas, classificadores semânticos, categorias gramaticais e localização geográfica. Neste trabalho apresentam-se os principais critérios aplicados no processo de lematização, que implica a atribuição de lemas e de categorias gramaticais às variantes, quer dizer, às unidades léxicas registradas nas fontes. Os lemas têm como finalidade agrupar as diferentes variantes flexivas, ortográficas ou fônicas que se encontram nas obras introduzidas na base, já, as categorias servem para unificar a diversidade de etiquetas e de informação gramatical que se encontra nelas. Além disso, apresentam-se as duas vias de ampliação do projeto: o estabelecimento de geossinônimos e a conexão dos geossinônimos e dos lemários galegos e portugueses.

Palavras-chave: Dialetologia. Lexicografia. Lematização.

* Instituto da Lingua Galega da Universidade de Santiago de Compostela. Santiago de Compostela, Galiza, Espanha. Contato: maria.alvarez.delagranja@usc.es.
** Programa de doutoramento Filoloxía Galega. Instituto da Lingua Galega da Universidade de Santiago de Compostela. Santiago de Compostela, Galiza, Espanha. Contato: marta.negro@usc.es.

[1] Este trabalho está situado no âmbito do projeto Tesouro do Léxico Patrimonial Galego e Português, que foi desenvolvido com financiamento do Ministerio de Ciencia e Innovación (Espanha) (FFI2009-12110) e da Fundação para a Ciência e a Tecnologia (Ministério de Ciência, Tecnologia e Ensino Superior de Portugal) (PTDC/CLE-LIN/102650/2008). A manutenção e ampliação do projeto é possível graças ao auxílio financeiro da Consellería de Cultura, Educación e Ordenación Universitaria da Xunta de Galicia a grupos de referência competitiva (GRC2013-040), cofinanciado parcialmente pelo FEDER, e ao convênio de colaboração entre esta mesma Consellería e a Universidade de Santiago de Compostela para o financiamento de diversas atividades de investigação do Instituto da Língua Galega.

ISBN 978-85-7846-344-1

Introdução

O *Tesouro do Léxico Patrimonial Galego e Portugués* (TLPGP) é um projeto que tem por objetivo integrar em um único banco de dados, de consulta livre através de Internet, material léxico dialetal do galego, do português de Portugal e do português do Brasil com indicação da sua distribuição geográfica. O TLPGP está coordenado pela professora Rosario Álvarez Blanco, do Instituto da Lingua Galega da Universidade de Santiago de Compostela (USC), e nele participam 18 universidades diferentes (além da USC, as Universidades portuguesas de Lisboa e Coimbra, assim como 18 universidades brasileiras). O projeto iniciou-se em 2007 e desde 2014 pode-se aceder livremente à aplicação de consulta em <http://ilg.usc.es/Tesouro/>, onde também é possível encontrar informação sobre as características fundamentais do TLPGP e sobre as diferentes equipes de trabalho.

O vocabulário compilado, em contínuo enriquecimento, extrai-se de fontes de diversos tipos: dicionários e glossários, atlas linguísticos ou mesmo obras «redigidas» que recolhem léxico dialetal e explicam o seu significado. Esses trabalhos podem estar já publicados, mas o TLPGP se nutre especialmente de obras inéditas de difícil acesso, em boa medida trabalhos acadêmicos (teses, trabalhos de mestrado...). Dado que um dos objetivos principais do projeto é mostrar a distribuição das diferentes formas através dos territórios galego, português e brasileiro, é condição indispensável para a integração dos itens lexicais na base de dados que estes materiais estejam localizados geograficamente. A aplicação de consulta oferece a correspondente informação, com representação cartográfica associada.

Neste trabalho mostramos as vias e os critérios aplicados para unificar as diferentes formas lexicais que se podem encontrar nas diversas fontes do TLPGP. Com esta finalidade, depois de apresentar brevemente o sistema utilizado para introduzir na base de dados a informação contida nas obras, justificamos a necessidade de atribuir lemas, apresentando os principais critérios usados para determinar as formas agrupáveis sob um mesmo lema e assinalando os critérios usados para selecionar este. A seguir apresentamos o sistema de atribuição de classes gramaticais e o tratamento das expressões complexas. Finalmente, assinalamos as vias de ampliação do projeto programadas atualmente.

Anais do III CIDS – Congresso Internacional de Dialetologia e Sociolinguística – Variedade, Atitudes Linguísticas e Ensino
Homenagem a Jacyra Andrade Mota e Suzana Alice Marcelino Cardoso
Londrina, Universidade Estadual de Londrina – 7 a 10 de outubro de 2014
ISBN 978-85-7846-344-1

O Tratamento da Informação no TLPGP

As obras selecionadas são editadas pelos membros do projeto para a sua integração na base de dados que constitui o TLPGP, conforme um protocolo previamente estabelecido (documento de trabalho "Tratamento dos materiais e estrutura da base de dados"). A informação registrada nos glossários, atlas e demais trabalhos é classificada em diversos campos para o seu posterior processamento (campos com o fundo cinzento no Quadro 1). Os editores introduzimos também outros dados que permitem organizar e unificar a informação (campos com o fundo branco). No Quadro 1, podemos verificar os diferentes campos empregados na base.

Quadro 1 – Os campos da base de dados do TLPGP

Introdução dos dados (por obra)															Tratamento dos dados			
															Por obra			
1	2	3	4	5	6	7	8	10	11	11b	12	13	15	20	9	16	17	6
variante	fonética	classe e categoria	definição	exemplos e refrães	comentários manipulador	página	secção	citação bibliográfica	código geográfico	existe tabela localidades	imagens	termo remissão	remissão a textos	criação registo	classificação semântica	lema de cada língua	classe e categoria	comentários manipulador

Deste jeito, por exemplo, às entradas *lancheira* e *lancho*, extraídas de Buescu (1961), corresponderiam, na base de dados do TLPGP, as que se mostram no Quadro 2[2].

[2] No Quadro 1 registram-se todos os campos que podem aparecer na base de dados, mas em cada uma das entradas que construímos só figurarão os que sejam pertinentes. Assim, no Quadro 2, não há coluna 2 nem 5, posto que as entradas de Buescu (1961) não têm informação fonética nem exemplos.

ISBN 978-85-7846-344-1

LADEIRA — *s. f.*, terreno inclinado.
LAJA — *s. f.*, rocha de superfície plana.
LANCHAL — *s. m.*, lugar onde há *lanchos*.
LANCHEIRA — *s. f.*, lugar onde há *lanchos*.
LANCHO — *s. m.*, penedo.
LAPA — *s. f.*, gruta.
LAPACHEIRO — *s. m.*, lamaçal.
LINDA — *s. f.*, limite, *malhão*.

MALHÃO — *s. m.*, marco que limita as propriedades.

Imagem 1 – Fragmento de Buescu (1961) em que figuram as formas *lancheira* e *lancho*

Quadro 2 – Tratamento no TLPGP das formas *lancheira* e *lancho* extraídas de Buescu (1961)

1	3	4	7	8	9	10	11	13	16	17
lancheira	s.f.	Lugar onde há lanchos.	320	A Terra. Acidentes e limites do terreno	2.2	Buescu 1961	0505	lancho	lancheira	v
lancho	s.m.	Penedo.	320	A TERRA. Acidentes e limites do terreno	2.2	Buescu 1961	0505	lancheira	lancho	sm

A Unificação de Variantes Através dos Lemas

No campo 1 (Quadros 1 e 2), é incluída a expressão conforme consta na fonte original. Essa forma é denominada de "variante". Salvo pequenas adaptações (por exemplo, substituição, se for o caso, de marcas fonéticas por caracteres tradicionais), e com o objetivo de manipular o menos possível as obras recolhidas no TLPGP, as variantes são mantidas em sua forma original.

Para compreender a necessidade de estabelecer lemas (campo 16 dos Quadros 1 e 2), é preciso levar em consideração a diversidade existente na forma de apresentação das variantes nas diferentes obras introduzidas. Freqüentemente

ISBN 978-85-7846-344-1

aparecem na forma de citação canônica: infinitivo nos verbos (*petiscar*), singular nos substantivos variáveis em número (*casa*; *abellón*), masculino singular nos adjetivos e nos substantivos variáveis em gênero e número (*afillado*; *casadeiro*). Todavia, não é infreqüente encontrá-las também em forma de lemas múltiplos, abreviados (*afillado, -a*) ou desenvolvidos (*casadeira, casadeiro*; *abellón*, *abellós*), ou mesmo em uma forma flexionada diferente da canônica (*petiscando*; *casas*; *afillada*; *casadeira*). Essas são as três variantes correspondentes a *casadeiro* que se registram no TLPGP:

> **casadeira** casad[ɛj]ra Mujer en edad de matrimonio. Taboada 1971:28. LEMA: casadeiro a.
>
> **casadeira, casadeiro** Soltera, soltero de unos 30 hasta 40 años. Schneider 1938:271. LEMA: casadeiro a.
>
> **casadeiro** casad[ɛ]iro adx. Que está en edade de casarse. Bravo 1984:64. LEMA: casadeiro a.

É preciso levar em conta, também, a variabilidade ortográfica que podemos encontrar em muitas formas nas distintas obras, sobretudo, se no momento de elaborarem estas os autores careciam de uma referência padrão. Assim, por exemplo, no TLPGP encontramos as variantes *lage* e *laje*, *ucha* e *hucha* ou *de cote* e *decote*.

A diversidade na apresentação formal e a diversidade ortográfica justificam por si a necessidade de estabelecer, na estrutura da base, algum campo unificador das variantes que permita realizar uma busca conjunta de todas elas. Esse campo é o denominado "lema" e, como indicamos, corresponde ao número 16 nos Quadros 1 e 2. Deste modo, os editores atribuímos a todas as variantes um lema normalizado de acordo com critérios de lematização previamente estabelecidos (documento de trabalho "Critérios de lematização"), em que a forma de citação escolhida se corresponde com a habitual na tradição lexicográfica galega e portuguesa (grosso modo, e sem entrar em detalhes, singular nas palavras variáveis em número, masculino nas variáveis em gênero, infinitivo nos verbos). Assim, por exemplo, às variantes *casadeira*; *casadeiro*, e *casadeira, casadeiro* atribuimos o lema masculino singular *casadeiro*, a *abellón* e a *abellón, abellós* o lema singular *abellón* ou às variantes *petiscar* e *petiscando* o lema em infinitivo *petiscar*.

De igual maneira, os lemas permitem unificar variantes ortográficas: as variantes portuguesas *lage* e *laje* unificam-se sob o lema *laje* ou as variantes galegas *de cote* e *decote* sob *decote*. O lema escolhido coincide com a forma padrão, portuguesa ou galega, segundo corresponda em cada caso. A este respeito, ressaltamos que os lemários galego e português são independentes, embora esteja previsto conectá-los no futuro.

ISBN 978-85-7846-344-1

Portanto, na aplicação, o usuário, além de poder realizar buscas pela variante concreta, oferecida na fonte original (por exemplo, por *lage* ou por *laje*), também pode procurar através do campo "lema", o que lhe dará como resultado o conjunto de todas as variantes a que se atribuiu o lema inserido na caixa de buscas (através de *laje* chega-se a *lage*, *laje* ou *lages*). Na representação cartográfica, iluminam-se conjuntamente as localidades associadas a essas variantes.

Sobre o assunto, consideramos que a unificação de variantes não deve restringir-se aos casos indicados, se quisermos evitar uns resultados demasiadamente atomizados. É evidente que o lugar em que se situa a fronteira entre "duas realizações de uma mesma unidade léxica" e "duas unidades léxicas distintas" é relativamente arbitrário, mas entendemos que variantes que mostram fenômenos fônicos não consolidados sempre historicamente (metáteses, harmonizações vocálicas...), como os que encontramos em *pedir* / *pidir*; *bailar* / *beilar*; *cadáver* / *cadavre* / *cadavle* ou *semana* / *somana*, devem estar disponíveis conjuntamente por meio de uma única busca, ou seja, devem constar sob o mesmo lema, respectivamente, as formas padrão *pedir*, *bailar*, *cadáver* e *semana*. Contrariamente, consideramos que itens com diferenças morfológicas, como as formas portuguesas *azinho e azinheira* ou as vozes galegas *apalleirar* e *empalleirar* ('fazer palheiros'), devem ter o seu lema diferenciado (*azinho*, *azinheira*; *apalleirar*, *empalleirar*).

Por outro lado, foi preciso trabalhar em um primeiro momento com critérios de lematização operativos, que não encerrassem uma casuística muito complexa, critérios que não nos obrigassem a tomar decisões específicas a cada passo ou a realizar um estudo pormenorizado de cada forma. Por tal motivo, optamos por fazer uma generalização a partir das decisões assinaladas e adotamos:

a) lematizar sob um mesmo lema "as formas de igual significado que se podem considerar variantes fônicas (seja por tratar-se de resultados distintos de um mesmo étimo, por determinada forma ser o resultado da evolução doutra ou por estarmos diante de processos fonéticos não consolidados historicamente, mas frequentes no discurso: harmonizações vocálicas, labializações...)", ("Critérios de lematização [textos portugueses]", p. 9). Assim, as formas galegas *camiño* e *camín*, que são resultados distintos de um mesmo étimo, unificam-se sob o lema padrão *camiño*, de igual modo, as variantes portuguesas *bexiga* e *buxiga*, que apresentam um fenômeno de alternância vocálica, agrupam-se sob o lema padrão português *bexiga*.

b) lematizar sob lemas diferentes, além das formas de igual significado que não partilham raiz (*pimpín* / *chincho*; *tornozelo* / *artelho*), "as formas de igual significado que, mesmo que partilhem a raiz, mostram diferenças morfológicas, quer porque uma procede da outra através de um processo morfológico sincrónico, quer porque provêm

ISBN 978-85-7846-344-1

de étimos com a mesma raiz, mas que se diferenciaram morfologicamente entre si num dado momento da história da língua" ("Critérios de lematização [textos portugueses]", p. 12). Os exemplos acima oferecidos (*azinho* / *azinheira*, *apalleirar* / *empalleirar*) ilustram as duas possibilidades assinaladas.

Vejamos a aplicação destes critérios em um exemplo concreto. Para designar o animal de nome científico *Talpa europaea,* encontramos no TLPGP as seguintes variantes galegas: *teipa*, *teipe*, *teupa*, *tiopa*, *topia*, *toupa*, *toupeira* e *cuvaterra*. As seis primeiras formas, que podem ser consideradas variantes fônicas, agrupam-se sob a forma padrão *toupa*. A variante *toupeira* não é unificável com as anteriores por possuir uma evidente diferença morfológica a respeito delas (sufixo *-eira*), apresentando, por isso, o seu próprio lema, coincidente com a variante: *toupeira*. Finalmente, *cuvaterra*, que possui raiz diferente à anterior (*toup-*), também figura por tal motivo com um lema diferenciado. O lema atribuído é a sua variante fônica *cavaterra*, forma que, embora não tenha caráter padrão nem esteja registrada no TLPGP, está recolhida em outros dicionários galegos[3], assim como em dicionários portugueses, neste caso sob a forma *cava-terra* (*Dicionário Priberam da Língua Portuguesa* 2008-2013, *Dicionário da Língua Portuguesa com Acordo Ortográfico* 2003-2015).

Os critérios gerais apresentados foram matizados ou complementados com alguns mais específicos, os quais não podemos mostrar com detalhes nesta oportunidade, mas que ilustraremos com algum exemplo concreto: assim, pelas dificuldades de discernir entre um fenômeno morfológico (presença do prefixo *a-*) ou fônico (aférese ou prótese da vogal *a-*), as formas em que a diferença se reduz à presença / ausência de um *a* inicial (*apandar* / *pandar*, *alustro* / *lustro*) agrupam-se sob um mesmo lema; também se recuperam conjuntamente as palavras que sofrem modificações formais por etimologias populares ou remotivações e aquelas de que partem: *vagabundo* / *vagamundo*, *mandarina* / *mondarina*, *teleférico* / *telesférico* etc.

Os critérios gerais fizeram-se extensíveis aos frequentes empréstimos do castelhano no galego[4], de tal jeito que as formas galegas e castelhanas que procedem

[3] Concretamente, em *Frampas, contribución al diccionario gallego*, de Eligio Rivas Quintas, s.v. Esta obra só se pode consultar através do *Dicionario de Dicionarios* (Santamarina 2006-2013).

[4] Durante vários séculos, o uso do galego restringiu-se a situações coloquiais e familiares, e as funções "altas" foram ocupadas pelo castelhano. Devido a este facto, os mecanismos de renovação léxica e terminológica da língua galega viram-se inibidos e muitas palavras do castelhano penetraram nela, quer porque em galego não existia uma forma própria para designar a realidade em questão (realidades novas, relacionadas frequentemente com o avanço da tecnologia), quer porque, existindo uma palavra galega com a mesma designação que a castelhana, se produziu uma substituição ou uma relegação da primeira, essencialmente por motivos de prestígio. Deste modo, ainda hoje se empregam no galego popular numerosos castelhanismos como *grifo*, *corbata, camilla, riñón, abuelo, iglesia, silla*... Os trabalhos sobre castelhanismos são muito numerosos. Podem consultar-se, entre outros, Chacón Calvar (2002), Dubert García (2005) ou Parga Valiña (2004).

ISBN 978-85-7846-344-1

do mesmo étimo se recuperam conjuntamente através de um mesmo lema (*avó* agrupa a forma de criação galega *avó* e o castelhanismo *abuelo*, ambos procedentes do lat. vulg. *AVIŎLUS). Pelo contrário, aquelas que partilham raiz, mas divergem morfologicamente, têm lemas distintos: a palavra de origem galega *muíño* e o castelhanismo *molinillo*, que incorpora à raiz o sufixo *-illo* lexicalizado, lematizam-se separadamente.

A Seleção dos Lemas

O critério geral adotado no momento de selecionar um lema para uma variante foi o seguinte: quando for possível, o lema deve coincidir com a forma padrão, com a forma referendada e priorizada pelos dicionários (com a galega para as variantes galegas e com a portuguesa para as variantes portuguesas, posto que, como já mencionamos, os lemários galegos e portugueses são independentes entre eles). Assim, por exemplo, as variantes fônicas do galego *caluba*, *caluga* e *culiga* têm como lema *caluga*, que é a forma normativa correspondente na citada língua.

Mas não podemos perder de vista o sentido do processo de lematização. O lema serve para agrupar variantes flexivas, ortográficas ou fônicas e ele próprio tem que coincidir com a forma que em cada caso se está lematizando ou bem ser uma variante flexiva, ortográfica ou fônica dela.

Em galego, a forma padrão para designar a ave de nome científico *Alauda arvensis* é *laverca*. No TLPGP, além de *laverca,* encontramos variantes fônicas, ortográficas e flexivas como *labiarca*, *laberca* e *labercas* que, segundo o indicado, lematizam-se sob a forma padrão *laverca*, mas também sob as formas *labarquela* e *levarquela*. Nenhuma dessas duas formas e nenhuma variante fônica ou ortográfica delas é uma palavra padrão, recolhida como tal nos dicionários galegos de referência. Apesar disso, não podemos lematizar *labarquela* e *levarquela* sob a forma padrão *laverca*, posto que o lema não pode apresentar diferenças morfológicas com respeito às variantes. Assim, atribuímos a *labarquela* e *levarquela* o lema não padrão *laverquela*, variante coerente com *laverca* e da que existem testemunhos galegos (SÁNCHEZ RODRÍGUEZ, 2000, p. 910).

Isso posto, inferimos que não será sempre possível encontrar um lema padrão no processo de lematização das variantes. Em primeiro lugar, porque estas, possuindo a mesma raiz que a forma normativa, podem apresentar diferenças morfológicas com respeito a ela, tal e como acabamos de verificar no exemplo anterior ou como sucede com *escachear*, não lematizável pela causa assinalada sob a forma padrão *cachear*. Mas também pode acontecer que as palavras recolhidas nas obras introduzidas não

ISBN 978-85-7846-344-1

estejam registradas nos dicionários sob nenhuma variante, nem fônica nem morfológica, como ocorre, por exemplo, com *investres* 'envolturas fetais', *arreleixar* 'tocar a morto' ou *rábade* 'comida de peixes'.

Para os casos em que o lema não possa ter caráter padrão, arbitramos outros critérios de seleção, a saber: o etimológico, a coerência na família de palavras e a proximidade às línguas próximas (sobretudo ao português do galego, e vice-versa). Vejamos um exemplo de aplicação simultânea dos três critérios:

Em um glossário de Goián (Tomiño-Galiza) (PÉREZ ALONSO, 1969, s.v.), recolhe-se a forma *trovexar* ('tronar'). No galego padrão não há nenhuma palavra de formação similar e as bases padrão sobre as que se poderia criar o derivado são *trebón* (que daria origem a **trebexar*) ou *torbón* (que daria origem a **torbexar*). Dado que não temos constância da existência de nenhum desses verbos e posto que a forma registrada em Goián coincide com uma palavra do português padrão, *trovejar*, consideramos conveniente lematizar com a variante registrada, *trobexar*, com uma ligeira mudança ortográfica: a grafia <b> vem exigida pela família de palavras (*trebón*, *torbón*), concordante com o étimo (lat. vulg. TURBO, -ÔNIS; vid. Corominas s.v. *turbar*)[5].

Caso desconheçamos a origem etimológica da variante e não sejamos conscientes da existência de outras formas da mesma família de palavras nem de equivalentes portugueses padrão, optamos por manter a forma tal qual a registramos. Seja como for, as variantes problemáticas etiquetam-se com uma marca de dúvida que nos permitirá localizá-las posteriormente, já que é muito possível que os materiais que vamos recolhendo no TLPGP nos possam ajudar a encontrar a família de palavras "perdida".

A Atribuição de Categoria Gramatical

A atribuição de lema às variantes vai acompanhada também da atribuição de uma categoria gramatical normalizada, de acordo com uma série de etiquetas estabelecidas pelos editores. Levemos em consideração que os autores das obras introduzidas no TLPGP não atribuem sempre categoria gramatical ao léxico recolhido e, quando o fazem, seguem critérios muito diversos, tanto no referente às marcas empregadas (por exemplo, *sm* ou simplemente *m* para os substantivos masculinos), como à quantidade de informação oferecida (por exemplo, para os verbos alguns

[5] Como é sabido, em galego a distinção entre <v> e <b> é meramente gráfica. Ambas as grafias representam o mesmo fonema, /b/, não existindo a labiodental sonora /v/.

ISBN 978-85-7846-344-1

autores assinalam simplesmente a categoria superior, mas outros indicam subcategorias como transitivo, intransitivo etc.). Além disso, não é estranho encontrar divergências entre uns e outros em relação à categoria atribuída, como se pode verificar neste exemplo:

> **arremar** v.i. 3. Encostar: *Arrima-te a mim*. {Póvoa de Atalaia} Martins 1954:407. LEMA: arrimar v.
>
> **arrumar** v.t. 1. Encostar. | Usa-se no Alandroal (R.L. IV, 59). Carreiro 1948:136. LEMA: arrimar v.

Deste modo, a atribuição de marcas categoriais por parte dos editores (campo 17 do Quadro 1) permite generalizar e unificar a informação gramatical das entradas da base. Deve-se levar em consideração, por outro lado, que a categoria atribuída (que figura ao final das entradas) se vincula sempre com os lemas, não com as variantes, e tem em conta o funcionamento das palavras na variedade padrão. Isso implica que, em um exemplo como o que segue, a categoria dos lemas coincida em ambas as entradas (LEMA: nariz sm, apesar do gênero da variante ser distinto em cada uma delas (masculino na primeira entrada, feminino na segunda)[6].

> **narís** s.m. (Nariz). Parte saínte da cara situada entre a boca e a fronte. Monteagudo 1998. LEMA: nariz sm.
>
> **nariz** [na'riθ] sf Parte saínte da cara, entre a fronte e a boca, que intervén na respiración e onde reside o sentido do olfacto. Louredo 2012. Vid. narices. LEMA: nariz sm.

Como se pode comprovar nos exemplos anteriores, na aplicação de consulta, o usuário encontrará tanto a categoria introduzida pelos editores no campo 17, associada ao lema, como a informação oferecida pelos autores das obras e associada às variantes, se esta informação existir (campo 3 do Quadro 1).

Os editores só oferecemos informação sobre subcategorias (masculino, feminino e plural [plurais lexicalizados]) no caso dos substantivos, mas não no dos verbos, posto que a indicação sobre o caráter intransitivo, transitivo ou pronominal deste tipo de unidades complicaria em excesso os processos de categorização e ofereceria numerosos problemas. Essas dificuldades, ainda que em menor medida, também se

[6] O gênero feminino de *nariz* que encontramos em alguns falantes de galego tem a sua origem na influência do castelhano, onde a palavra é feminina.

ISBN 978-85-7846-344-1

podem apresentar para os substantivos. Como já assinalamos, alguns autores não oferecem informação categorial, resultando, em algum caso, na impossibilidade de saber qual é o gênero da variante recolhida, como sucede, por exemplo, com a voz galega *investres* mencionada na epígrafe anterior. Por tal motivo, arbitramos a etiqueta *des* (desconhecemos género) para substantivos e locuções substantivas (*s des*, *loc s des*).

Além de uma listagem de abreviaturas, os editores contamos com outras indicações sobre o processo de lematização e de atribuição de categoria gramatical que têm por objetivo tornar coerente o tratamento dos materiais entre os diferentes membros que trabalham no TLPGP ("Critérios de lematização [textos portugueses]", p. 3-8). Não podemos apresentá-las pormenorizadamente neste artigo, mas escolhemos estes dois pontos como ilustração do tipo de normas a que nos referimos:

> 1.3. As formas de **particípio**, tanto as regulares como as irregulares, lematizar-se-ão no infinitivo. Exceptuam-se aqueles casos em que o registo a lematizar corresponda claramente a um uso "consagrado" como adjectivo ou substantivo: *atrasado* 'tolo [pessoa]', *cozido* 'prato típico da cozinha portuguesa', etc).
>
> 2.6. Lematizar-se-ão separadamente os **heterónimos** (*boi* | *vaca*, *cavalo* | *égua*, *cão* | *cadela*) e também os femininos que se criam mediante **sufixos derivacionais**: *abade* | *abadessa*, *conde* | *condessa*, *actor* | *actriz*, *herói* | *heroína*, *galo* | *galinha*. Assim sendo, será atribuído o lema que corresponda à forma (masculina ou feminina) que apareça no registo. Se neste aparecerem as duas formas, então teremos de duplicá-lo, para poder atribuir ambos os lemas: *galo* ⇒ *galo* sm; *galinha* ⇒ *galinha* sf; *galo*, *galinha* ⇒ *galo* sm | *galinha* sf.

O Tratamento das Expressões Complexas

Os editores do TLPGP consideramos que as locuções e os compostos sintagmáticos, com independência de que nas fontes originais figurem como sublemas dentro do artigo lexicográfico de algum dos seus componentes, devem ter entrada independente na base de dados, por serem unidades léxicas de pleno direito.

Por tal motivo, caso ocorra a circunstância mencionada no parágrafo anterior, as expressões serão extraídas dos seus lemas ordenadores para conferir-lhes *status* independente. No lema ordenador, deixamos uma pegada da existência da expressão complexa mediante a criação de uma remissão. Desse jeito, a uma entrada como esta, tirada de Monteagudo (1998),

ISBN 978-85-7846-344-1

> **bata** *sf* Prenda de vestir frouxa e aberta por diante ou polo lombo con botóns. BATA DE CASA *loc. subs* Prenda de vestir empregada para estar na casa; sobre todo ponse ó levantarse da cama.

correspondem três entradas na base, na qual a segunda tem caráter remisivo:

> **bata** s.f. Prenda de vestir frouxa e aberta por diante ou polo lombo con botóns. Monteagudo 1998. LEMA: bata sf.
>
> **bata** Monteagudo 1998. Vid. bata de casa. LEMA: bata sf.
>
> **bata de casa** loc. subs Prenda de vestir empregada para estar na casa; sobre todo ponse ó levantarse da cama. Monteagudo 1998. LEMA: bata de casa loc sf.

Caso a expressão complexa já possua na fonte original entrada independente, nós simplesmente mantemos esse *status*.

A atribuição de lema próprio às expressões complexas implica a necessidade de atribuir-lhes também categoria gramatical. Dessa forma, dentre a listagem de etiquetas que utilizamos, podemos encontrar as correspondentes a locuções de diversos tipos, equivalentes a praticamente todas as classes de palavras: verbais, substantivas masculinas, substantivas femininas, adjetivas, adverbiais etc. Para não complicar em extremo o labor de edição e simplificar a classificação, os compostos sintagmáticos de caráter nominal (*dente do siso*, *bata de casa*) são etiquetados como locuções substantivas.

No entanto, é muito habitual que os dicionários e vocabulários lematizem ou sublematizem como expressões complexas combinações frequentes que, em realidade, não devem ser consideradas locuções nem compostos sintagmáticos. Encontramos nas fontes originais lemas ou sublemas como *andar às cavaleiras* ou *segar a herba*, que não constituem expressões fixadas, pois não são unidades léxicas memorizadas como um todo, senão a soma de várias unidades que simplesmente coocorrem freqüentemente (como demonstra o fato de poder combinar *às cavaleiras* com outros verbos diferentes a *andar*[7], ou de poder modificar livremente

[7] Para ilustrar, oferecemos três exemplos extraídos do *Corpus do Português: 45 million words, 1300s-1900s*: "Os escravos *correram* com Calpúmio *às cavaleiras*, e era como se transportassem a própria morte", "Pesarosos, senão amedrontados das responsabilidades, deviam estar aqueles sustentáculos da ordem, todos eles de lombos maduros de *carregar* os filhos *às cavaleiras*"; "...aquela Joaninha com quem eu andava ao colo, que *trazia às cavaleiras*, que me fazia ser tão doido e tão criança como ela".

ISBN 978-85-7846-344-1

o substantivo *herba*[8], ou substituir verbo e complemento por outras palavras sem que mude o significado do outro elemento). Os editores devemos diferenciar entre as unidades lexicalizadas e aquelas que não o estão, pois o tratamento em cada caso é distinto. No primeiro, como já indicamos, outorgamos um lema complexo à expressão (assim como a etiqueta categorial *loc* seguida da especificação correspondente). No segundo, outorgamos um lema a cada uma das unidades léxicas que constituem a combinação com uma etiqueta (nos casos acima citados, os lemas são *andar* v, *às cavaleiras* loc adv e *segar* v, *herba* sf).

Caso não tivéssemos estabelecido esse procedimento, e continuando com o exemplo anterior, as variantes que figuram no TLPGP *às cavaleiras*, *levar às cavaleiras* e *andar às cavaleiras* teriam lemas distintos e não seriam recuperáveis conjuntamente. De tal modo, a informação sobre a distribuição da locução adverbial *às cavaleiras* seria mais difícil de recuperar.

Cabo

O TLPGP sistematiza e ordena por meio dos lemas uma grande quantidade de formas léxicas, entre as quais é necessário destacar a presença de numerosas unidades pluriverbais e de muitas palavras, não registradas em dicionários comuns, que figuravam dispersas em diferentes obras dialetais, quase sempre de difícil acesso. No futuro está previsto ampliar as utilidades do projeto por meio de duas vias:

1) Através de um segundo processo de lematização consistente na vinculação dos lemas sinônimos a um único geossinônimo de referência, de tal modo que o usuário possa conhecer quais são as diferentes formas de expressar o mesmo conceito por meio de uma única busca: o geossinônimo português *arco-da-velha* agrupará os lemas portugueses *arco-da-velha, arco-celeste, arco-da-aliança, arco-íris, arco-da-Santíssima-Trindade, arco-de-nossa-senhora*... com as variantes correspondentes; o geossinônimo galego *toupa* dará acesso aos lemas galego *toupa*, *toupeira* e *cavaterra* com todas as suas variantes.

2) Através da conexão dos lemários e dos geossinônimos galegos e portugueses.

Quando essas duas tarefas estiverem ultimadas, por meio do TLPGP, os usuários poderão obter, mediante uma única busca por um geossinônimo, as diferentes formas

[8] Como se pode observar nesses dois exemplos tirados do *Tesouro Informatizado da Lingua Galega*: "A tradición conta que un home andaba a segá-*la herba dun prado* coa gadaña", "...o día foi longo e houbo moito que bracear para segar *a herba toda*".

ISBN 978-85-7846-344-1

de expressão ou variantes existentes para o conceito correspondente nos territórios galego, português e brasileiro, agrupadas por lemas. Seguiremos a trabalhar, pois, com o objetivo de ampliar as utilidades do TLPGP para os dialetólogos, os lexicólogos e para todos aqueles interessados no conhecimento do vocabulário dialetal galego e português, e com a intenção de oferecer um panorama cada vez mais enriquecedor e completo do nosso tesouro léxico.

Referências

ÁLVAREZ, R. (Coord). **Tesouro do léxico patrimonial galego e portugués.** Santiago de Compostela: Instituto da Lingua Galega. Disponível em: <http://ilg.usc.es/Tesouro>. Acesso em: 28 jan. 2015.

BUESCU, M. L. C. **Monsanto.** Etnografia e linguagem. Lisboa: Centro de Estudos Filológicos, 1961.

CHACÓN CALVAR, R. Lenguas en contacto e interferencias. **Revista de lenguas y literaturas catalana, gallega y vasca**, Madrid, v. 8, p. 119-137, 2002. Disponível em: <http://migre.me/s3RZ2>. Acesso em: 28 jan. 2015.

COROMINAS, J.; PASCUAL, J. A. **Diccionario crítico etimológico castellano e hispánico**. Madrid: Gredos, 1991-1997.

DAVIES, M.; FERREIRA, M. **Corpus do Português:** 45 million words, 1300s-1900s. 2006-. Disponível em: <http://migre.me/s3S1A>. Acesso em: 28 jan. 2015.

DICIONÁRIO PRIBERAM da Língua Portuguesa [em linha]. 2008-2013. Disponível em: <http://migre.me/s3S22>. Acesso em: 28 jan. 2015.

DICIONÁRIO DA LÍNGUA PORTUGUESA com Acordo Ortográfico [em linha]. Porto: Porto, 2003-2015. Disponível em: <http://migre.me/s3S2E>. Acesso em: 28 jan. 2015.

DUBERT GARCÍA, F. Interferencias del castellano en el gallego popular. **Bulletin of Hispanic Studies**, Liverpool, v. 82, n. 3, p. 271-291, jun. 2005.

MONTEAGUDO CABALEIRO, M. T. **Contribución ó estudio do léxico do concello de Redondela**. 1998. Memória de Mestrado - Universidade de Santiago de Compostela, Santiago de Compostela.

ISBN 978-85-7846-344-1

PARGA VALIÑA, M. A interferencia lingüística no galego oral. In: ÁLVAREZ BLANCO, R.; FERNÁNDEZ REI, F.; SANTAMARINA, A. (Eds.). **A Lingua Galega:** historia e actualidade. Actas do I Congreso Internacional (Santiago de Compostela, 16-20 de setembro de 1996). Santiago de Compostela: Consello da Cultura Galega, 2004. v. I, p. 547-558. Disponível em: <http://migre.me/s3S3S>. Acesso em: 28 jan. 2015.

PÉREZ ALONSO, M. J. **Vocabulario de Goián**. 1969. Memória de Mestrado - Universidade de Santiago de Compostela, Santiago de Compostela.

SÁNCHEZ RODRÍGUEZ, H. Léxico da parróquia de Seteventos. In: RODRÍGUEZ, J. L. (Ed.). **Estudos dedicados a Ricardo Carvalho Calero**. Santiago de Compostela: Parlamento de Galicia; Universidade de Santiago de Compostela, 2000. v. I, p. 903-924.

SANTAMARINA, A. (Coord.). **Dicionario de dicionarios.** Corpus lexicográfico da lingua galega. 2006-2013. Santiago de Compostela: Instituto da Lingua Galega; Vigo: Universidade de Vigo. Disponível em: <http://migre.me/s3S5t>. Acesso em: 28 jan. 2015.

SANTAMARINA, A. (Coord.). **Tesouro informatizado da lingua galega**. 2015. Santiago de Compostela: Instituto da Lingua Galega. Disponível em: <http://ilg.usc.es/TILG/>. Acesso em: 28 jan. 2015.