CORTEGAL, Corpus de textos galegos escritos por estudantes no ámbito académico. Deseño do corpus e caracterización dos textos

María Álvarez de la Granja
Instituto da Lingua Galega-Universidade de Santiago de Compostela

Reyes Rodríguez Rodríguez
Instituto da Lingua Galega-Universidade de Santiago de Compostela

1. Introdución

CORTEGAL, Corpus de textos galegos escritos por estudantes no ámbito académico, é un corpus anotado de textos escritos en lingua galega por alumnado de Galicia nas probas de acceso á universidade. O obxectivo fundamental deste corpus, elaborado no Instituto da Lingua Galega da Universidade de Santiago de Compostela e accesible en http://ilg.usc.gal/cortegal/, é coñecer algunhas características da escrita académica do estudantado galego ao remate da educación secundaria. Particularmente, a partir da súa análise poderán detectarse as principais eivas e carencias do alumnado na expresión escrita en lingua galega e artellar *a posteriori* estratexias e recursos didácticos focalizados nos aspectos máis problemáticos. Tamén confiamos en que a análise do corpus sirva para facilitar e promover outros estudos sobre a escrita en lingua galega do alumnado, máis alá das diverxencias do estándar e particularmente no ámbito discursivo. Así mesmo, un terceiro obxectivo é converter *CORTEGAL* nunha ferramenta didáctica de aplicación directa nas aulas, de tal xeito que o alumnado traballe con textos escritos por outros/outras estudantes, moi en especial no marco da aprendizaxe guiada por datos («data driven learning») (Gilquin e Granger 2022; McEnery e Richard 2010; Römer 2011; Timmis 2015; Vázquez Rozas e Blanco 2022).

A finalidade deste capítulo é dobre. Por un lado, pretendemos realizar unha presentación xeral da metodoloxía empregada na elaboración de *CORTEGAL*, tanto no relativo á confección da mostra como no que ten que ver coa transcrición, anotación e estandarización das redaccións que conforman o corpus. En primeiro lugar, en §2.1, describiremos os criterios de selección dos textos e ofreceremos unha caracterización xeral destes desde o punto de vista cualitativo (tipo de textos, temas, cualificacións...). A seguir, en §2.2, amosaremos, en liñas xerais, o proceso de tratamento dos textos levado a cabo na plataforma *TEITOK*, dando conta das anotacións que conteñen. A continuación, en §2.3, abordaremos con máis detalle a súa transcrición e en §2.4 a asignación de metadatos. A seguir, en §2.5 deterémonos na estandarización e codificación das formas non estándar e en §2.6 trataremos a asignación de lema e categoría gramatical. Finalmente, en §2.7 ofreceremos os trazos xerais da anotación dos conectores que serven para vincular enunciados. Esta presentación servirá como marco xeral para situar e entender adecuadamente os datos que se ofrecerán nas seguintes contribucións da obra e na segunda parte deste capítulo.

Nesta segunda parte presentaremos os datos cuantitativos globais do corpus, atendendo en §3.1 á caracterización xeral dos textos (número de palabras e lemas, media de parágrafos por texto, media de enunciados por parágrafo...), en §3.2 ao número de *tokens* e á súa distribución en clases de palabras, en §3.3 ás diferentes desviacións do estándar académico detectadas nos seis niveis analizados (ortográfico, morfolóxico, léxico, gramatical [sintáctico], semántico e discursivo) e finalmente, en §3.4, á presenza dos conectores entre enunciados. Finalmente, ofreceremos unhas conclusións en que se recollerá unha caracterización global dos textos que conforman o corpus atendendo á análise realizada previamente e unhas reflexións finais sobre *CORTEGAL*, a súa elaboración e o seu interese de cara á mellora da destreza da escritura en lingua galega.

2. Metodoloxía de elaboración de *CORTEGAL*

2.1. Os textos e a mostra

CORTEGAL é un corpus conformado por 1000 textos manuscritos redactados no ano 2017 por estudantes de Galicia no marco da proba de acceso á universidade denominada ABAU, «Avaliación do Bacharelato para o acceso á Universidade». Máis concretamente, os textos corresponden á proba de comentario que o alumnado debe realizar no exame de Lingua Galega e Literatura. Neste exame solicítase a redacción dun texto de carácter

argumentativo, de entre 200 e 250 palabras, sobre un tema determinado, vinculado cun texto previo. As redaccións, proporcionadas pola Comisión Interuniversitaria de Galicia (CIUG) cos correspondentes permisos para a elaboración do corpus, corresponden tanto á convocatoria de xuño como á de setembro do curso 2016-2017. Presentáronse ao exame un total de 8669 estudantes en xuño e 1197 en setembro, de tal modo que a mostra de *CORTEGAL* supón un 9,87% do total de exames presentados.

En cada unha destas convocatorias propóñense dous modelos de exame diferentes dos que o estudantado debe escoller un. O corpus ofrece redaccións correspondentes ás catro opcións posibles para a pregunta de comentario, que son as que se indican a seguir:

Xuño 2017

Opción A (texto inicial de Fran Alonso en *Dorna* 27, 2001)

> Nos últimos anos a gastronomía e a cociña acadaron moita popularidade. Redacta un texto expoñendo a túa opinión sobre este fenómeno: as súas causas, o que ten de moda pasaxeira ou de cambio cultural máis duradeiro...

Opción B (texto inicial de J. Luís Sucasas en *Vieiros*, 2009)

> Redacta un texto sobre a importancia que teñen o consumo e a produción (ou o consumismo e a produtividade) no noso modo de vida actual.

Setembro 2017

Opción A (texto inicial de Xavier Quiroga de *Zapatillas rotas*, 2014)

> Expón, de maneira argumentada, a túa opinión persoal sobre o problema que reflicte o texto e, en xeral, sobre este tipo de conflitos familiares entre pais e fillos adolescentes.

Opción B (texto inicial de Mercedes Queixas, en *Palavra Comum*, 09/10/2015)

> A autora móstrase crítica co feito de que a infancia e a mocidade soñe con ser futbolista ou modelo moi maioritariamente (liña 10). Redacta un texto expoñendo de maneira argumentada o teu acordo ou desacordo co seu punto de vista.

Os centros educativos aos que pertence o alumnado que realiza as probas ABAU están asignados a 26 Comisións Delegadas (en diante CD), de tal xeito que cada CD abrangue unha zona xeográfica ampla que recolle centros tanto públicos como privados e con alumnado de procedencia diversa (urbana, periurbana, vila ou rural). A listaxe de CD cos centros educativos asignados no curso 2016-2017 pode consultarse na páxina do corpus. O número de textos da mostra de *CORTEGAL* é proporcional á cifra total de exames por CD e convocatoria (xuño e setembro), con dúas puntualizacións: por un lado, eliminároнse os 29 exames correspondentes á CD 25, que recolle probas de alumnado con necesidades específicas procedente de toda Galicia e non dunha zona xeograficamente restrinxida; por outro lado, aínda que a

distribución real de exames entre xuño e setembro é de 87,9%-12,1%, na mostra de *CORTEGAL* a distribución mudouse lixeiramente (89,8%-10,2%), para reducir o peso do alumnado que realizou o exame tanto en xuño como en setembro (ao suspender na primeira convocatoria). En Álvarez de la Granja (2018, p. 58) explícase con máis detalle o procedemento seguido para levar a cabo esta modificación, que tivo en conta o número de suspensos na primeira convocatoria. A distribución exacta de exames por convocatoria e temática na mostra de *CORTEGAL* é a que se ofrece na Táboa 1:

Táboa 1. Distribución de exames por convocatoria e temática na mostra de *CORTEGAL*

Convocatoria	**Tema**	**Número de textos da mostra**	**Porcentaxe sobre o total de textos**
Xuño	A gastronomía	449	44,9%
	Consumo e produción	449	44,9%
Total xuño		*898*	*89,8%*
Setembro	Conflitos familiares	51	5,1%
	Os referentes da mocidade	51	5,1%
Total setembro		*102*	*10,2%*
Total		***1000***	***100%***

O reparto por temas non se fixo atendendo ás escollas do alumnado, senón que, tal e como se pode ver na táboa anterior, se procurou unha distribución equitativa entre as dúas opcións de cada convocatoria, para deste xeito asegurar a maior variedade posible nos textos e, daquela, no léxico. Máis alá desta circunstancia, a selección de textos para cada comisión foi aleatoria.

A nota media dos exames de Lingua Galega e Literatura de *CORTEGAL* é de 5,90 sobre 10 (cunha desviación estándar de 1,72) e da proba concreta de comentario de 6,75 sobre 10[1] (cunha desviación estándar de 1,93), co reparto entre xuño e setembro que se amosa na Táboa 2, onde se perciben diferenzas importantes entre as dúas convocatorias:

[1] As notas do comentario realízanse en realidade sobre 3 puntos, pero están convertidas aquí a base 10.

Táboa 2. Notas numéricas da mostra de *CORTEGAL*

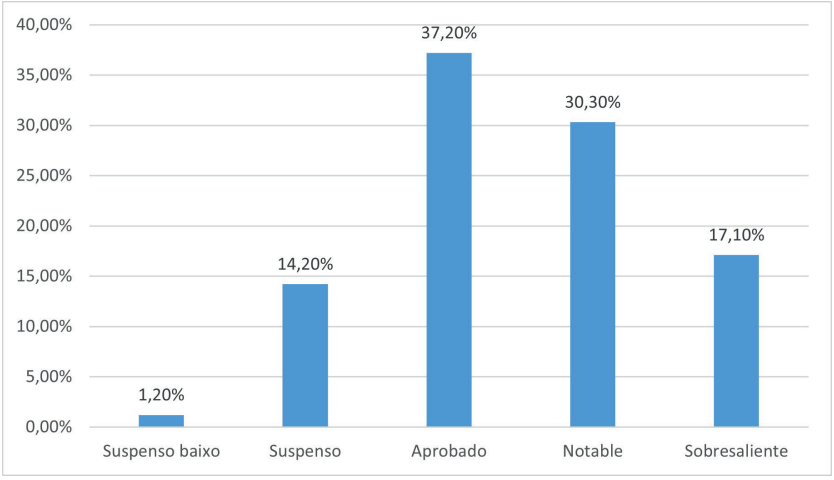
Convocatoria	Exame	Desviación estándar	Comentario	Desviación estándar
Xuño	6,09	1,65	6,93	1,87
Setembro	4,26	1,43	5,21	1,75
Xuño e setembro	5,90	1,72	6,75	1,93

A nota media do conxunto dos exames de Lingua Galega e Literatura das probas ABAU foi de 5,77 en xuño e de 4,17 en setembro. Vemos, pois, que as medias son bastante próximas ás da mostra, cunha diferenza de 0,32 en xuño e de 0,09 en setembro, sendo sempre máis alta a media dos textos do corpus. Carecemos de datos sobre os resultados da pregunta de comentario para realizar unha comparativa similar.

O rango de notas dos textos de *CORTEGAL* sitúase no exame de xuño entre 1 (nun único exame) e 10 (en seis exames) e no exame de setembro entre 0 e 8 nunha única proba para cada cualificación. No caso do comentario, o rango vai desde 0 nun só exame ata 10 en 45 exercicios en xuño, e dende 0 nun dos comentarios ata 9,33 noutro en setembro.

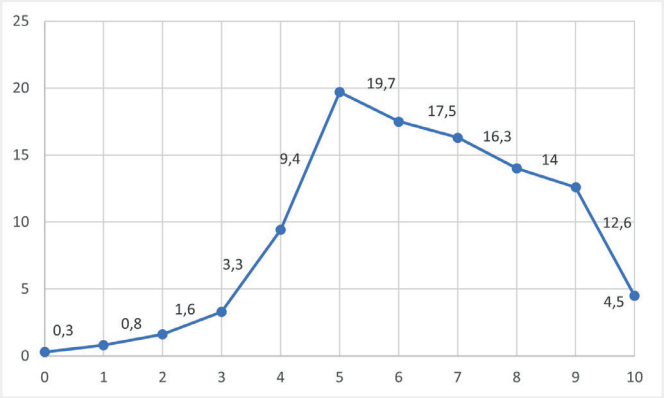
Na Gráfica 1 recóllese a distribución das cualificacións nos textos de *CORTEGAL*, onde Suspenso baixo abrangue desde 0 ata 2,4, Suspenso desde 2,5 ata 4,9, Aprobado desde 5 ata 6,9, Notable desde 7 ata 8,9 e Sobresaliente desde 9 ata 10. A cualificación máis frecuente é Aprobado, cun 37,20% dos exercicios. O 84,60% do alumnado superou a pregunta, mentres que o restante 15,40% obtivo unha cualificación inferior a 5.

Gráfica 1. Cualificacións dos textos de *CORTEGAL*



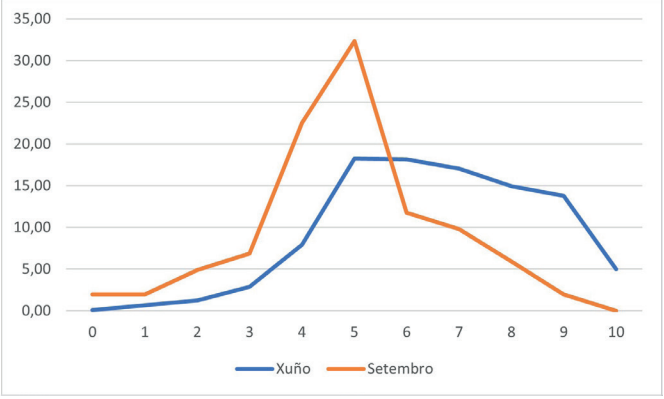
Na seguinte gráfica de dispersión (Gráfica 2), onde as cifras indican a porcentaxe de textos que obtiveron unha cualificación situada en cada rango numérico (0-0,9; 1-1,9 etc.), percíbese con máis detalle a distribución das notas. Comprobamos como o rango con maior número de exames é o situado entre 5 e 5,9, aínda que a nota concreta máis repetida, ou moda, é 6,67, moi próxima da media.

Gráfica 2. Distribución das cualificacións numéricas nos textos de *CORTEGAL*



Mais se se afonda en cada unha das convocatorias a través da Gráfica 3, compróbase como en xuño se debuxa unha liña relativamente estable dende a nota 5 cara adiante, en que se concentra máis do 87% do alumnado, mentres que en setembro, do 62% de aprobados, máis da metade pertencen ao rango 5-5,9 e a partir deste, cae dous terzos a cifra de cualificacións no rango 6-6,9, seguido dun descenso considerable a medida que se incrementa a cualificación.

Gráfica 3. Distribución das cualificacións numéricas dos textos de *CORTEGAL* en xuño e setembro



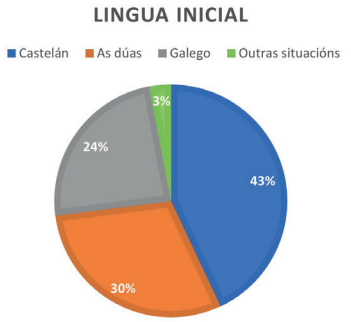
Malia estas diferenzas, compróbase como as dúas convocatorias coinciden en que a marca máis elevada se sitúa no rango 5-5,9 (aínda que en xuño a diferenza é mínima en relación co seguinte rango) e en que hai máis textos aprobados ca suspensos.

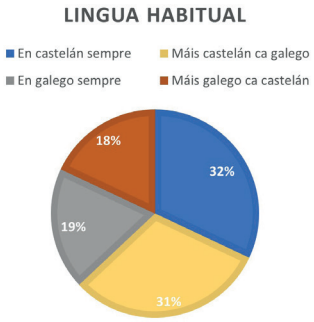
As cualificacións obtidas no exame de Lingua Galega e Literatura e na proba de comentario son uns dos poucos datos «extratextuais» que posuímos sobre as redaccións. Precisamente, un dos principais problemas implícitos no emprego de textos das probas ABAU é o feito de carecermos de datos sociolingüísticos sobre o alumnado que os redactou. Os exames son totalmente anónimos e non inclúen información de ningún tipo sobre xénero, lingua habitual, lingua inicial, tipo de residencia, nivel sociocultural etc. O único dato que posuímos corresponde, como indicamos, á CD ao que está asignado o centro educativo (este último tamén é descoñecido), de tal xeito que podemos atribuír a cada texto unha zona xeográfica ampla correspondente ao centro de ensino ao que pertence a/o estudante. Sabemos, en calquera caso, que se trata na súa maior parte de alumnado que acaba de rematar 2º de Bacharelato, maioritariamente de 17 ou 18 anos, e que ten o galego como L1 ou L2. Cómpre sinalar, de todos os xeitos, que o galego é lingua instrumental empregada na impartición de diversas materias ao longo de todos os cursos da educación primaria e da educación secundaria, así como obxecto de aprendizaxe nunha materia específica tamén en todos os cursos destes dous niveis educativos[2]. De acordo co indicado no artigo 14 da Lei 3/1983 de Normalización Lingüística, o estudantado, ao remate do ensino obrigatorio, deberá coñecer o galego, nos niveis oral e escrito, «en igualdade co castelán».

Dado que carecemos de datos sobre a lingua inicial e habitual das persoas autoras dos textos de *CORTEGAL*, podemos presentar a información ofrecida polo Instituto Galego de Estatística (IGE) (2019) na última «Enquisa estrutural a fogares», realizada en 2018, un ano despois de que fosen escritos os textos de *CORTEGAL*. Amosamos, na Gráfica 4 e na Gráfica 5, os datos relativos á franxa de idade en que se incluirían as persoas autoras das redaccións (entre 15 e 29 anos).

[2] O alumnado procedente doutras comunidades ou doutros países que se incorpore ao sistema educativo galego pode solicitar a exención da materia de lingua galega durante un máximo de dous anos consecutivos (https://www.xunta.gal/dog/Publicados/2014/20140219/AnuncioG0164-120214-0003_gl.html). O estudandanto que, por este motivo, non curse tal materia nun ou nos dous cursos de Bacharelato está exento da realización do correspondente exame nas probas ABAU (https://www.xunta.gal/dog/Publicados/2011/20110404/AnuncioEF0A_es.html; https://www.boe.es/buscar/pdf/2016/BOE-A-2016-7337-consolidado.pdf). Un total de 50 alumnos matriculados nas ABAU no curso 2016-2017 non realizaron o exame de Lingua Galega e Literatura.

Gráficas 4 e 5. Lingua inicial e lingua habitual da mocidade galega (15-29 anos) (Fonte: IGE, 2019)





Como se pode observar, a lingua inicial maioritaria é o castelán (43%), pero un 54% das persoas enquisadas indican que a súa lingua inicial é ben o galego, ben o galego e o castelán. Con respecto á lingua habitual, un 63% di falar só castelán ou máis castelán ca galego, mentres que un 37% indica que só fala galego ou que fala máis galego ca castelán. Evidentemente, a extrapolación destas porcentaxes a *CORTEGAL* cómpre facela con cautela, dado que a franxa de idade con que traballa o IGE (15-29 anos) é moito máis ampla ca a que corresponde, en liñas xerais, ao alumnado do corpus (17-18 anos).

No momento de seleccionar os textos das probas ABAU, eramos conscientes de que estes presentaban algunhas desvantaxes: a xa comentada imposibilidade de ter información sociolingüística sobre as persoas autoras das redaccións, a necesidade de transcribilas manualmente e a tensión derivada da importancia da proba e das restricións de tempo con que se elaboraron estas, que determinaron probablemente que o resultado non fose o mellor dos posibles (González Álvarez, 1990, p. 209). Con todo, o feito de traballar con textos das probas ABAU ofrece tamén varias vantaxes, tal e como se indica en Álvarez de la Granja (2018, p. 57). Entre elas, a dispoñibilidade inmediata dos textos, a súa homoxeneidade, así como a garantía de que a/o estudante escribe a redacción con seriedade, procurando utilizar sempre a variedade estándar (posto que calquera diverxencia pode ser obxecto de sanción na cualificación). Ademais, o feito de traballar con textos manuscritos permite, como veremos en §2.3, ofrecer información sobre as formas riscadas e as engadidas *a posteriori*, co que resulta posible rastrexar o proceso compositivo do alumnado. Finalmente, e de maneira especial, os textos das ABAU, pola súa transcendencia, teñen un valor engadido que dota o corpus, e os resultados que se extraian da súa análise, dunha especial relevancia no marco da sociedade galega.

O conxunto dos 1000 textos é accesible a través da pestana «Textos» da páxina do corpus (http://ilg.usc.gal/cortegal/index.php?action=texts). De igual xeito, se non se escribe nada na caixa do buscador e se preme en «Buscar» obtense tamén a listaxe de todas as redaccións, desta volta cos seus metadatos (*vid.* §2.4).

2.2. Tratamento dos textos en *TEITOK*

Os textos de *CORTEGAL* están transcritos, *tokenizados* e anotados na plataforma *TEITOK*, unha plataforma utilizada na creación de diferentes corpus, especialmente, aínda que non só, corpus históricos e de aprendentes, e de gran interese para o tratamento de textos manuscritos pola posibilidade de combinar anotacións lingüísticas (lema, categoría morfosintáctica...) e textuais (fragmentos borrados, cursivas...) (Janssen, 2016, p. 4037)[3].

> TEITOK is a framework for creating, maintaining, and publishing annotated corpora. It is a web-based environment written mostly in a combination of PHP and Javascript. In TEITOK, a corpus consists of a collection of XML files, each in the Text Encoding Initiative (TEI) format (...), with a slightly modified tokenization system (...). The system makes it easy to display each XML file (...), edit metadata (...) and individual tokens (...), and search through the corpus (...) (Janssen, 2016, p. 4037).

O proceso de traballo seguido cos diferentes textos que conforman o corpus *CORTEGAL* é o que se indica a seguir. Todas as tarefas, agás a primeira, foron realizadas en *TEITOK*:

1) Escaneado dos 1000 textos.
2) Creación dun arquivo XML coa transcrición manual do texto e etiquetas TEI que identifican a súa estrutura e algúns elementos concretos dos orixinais (formas riscadas, engadidas *a posteriori*...) (*vid.* §2.3).
3) Revisión das transcricións.
4) Introdución dos datos fixos da cabeceira e dalgúns datos variables (título, CD, cualificacións, convocatoria, tema, responsables da transcrición e revisión) (*vid.* §2.4).
5) *Tokenización*. Cada *token*, incluíndo signos de puntuación, é representado por un elemento <tok>. Debe terse en conta a este respecto que, no caso das contraccións e das unidades formadas por verbo e pronomes enclíticos, se emprega un sistema mixto, de tal xeito que toda a unidade é

[3] En http://teitok.corpuswiki.org/ pódese atopar unha listaxe de proxectos elaborados nesta plataforma, así como unha guía para o seu emprego.

un elemento <tok> conformado por varios *dtokens* (<dtok>) (*vid. infra* exemplo 2):

> One of the eternal problems in annotated corpora is how to handle contractions: whether to treat them as one token or two. That is why TEITOK takes a mixed approach: the <tok> elements are roughly speaking orthographic words. In the case of contractions, two or more grammatical words are inserted as children of the <tok>, called <dtok>. In this manner, it is possible to associate POS and lemma to the grammatical words, but associate the normalized orthography to the orthographic word (Janssen, 2016, p. 4038).

6) Estandarización e codificación das formas non estándar en seis niveis lingüísticos (*vid.* §2.5).
7) Lematización automática mediante *FreeLing*, con asignación de lema e categoría gramatical, e revisión manual posterior (*vid.* §2.6).
8) Asignación manual de lemas e categorías gramaticais orixinais (*vid.* §2.6).
9) Anotación dos conectores que vinculan enunciados (*vid.* §2.7).
10) Introdución da información restante na cabeceira dos textos (datos cuantitativos sobre os textos e responsables do proceso de anotación e lematización) (*vid.* §2.4).
11) Revisión da estandarización e codificación.
12) Revisión xeral do corpus.

No arquivo XML creado para cada texto, as anotacións lingüísticas son representadas normalmente como atributos dos diferentes *tokens* (para excepcións, *vid.* §2.5.4). No exemplo (1) que ofrecemos a seguir, a forma escrita polo estudante, *recetas* (*token* 137 do texto), recibe as seguintes anotacións: a forma lexicamente estandarizada, *receitas*, o lema orixinal *receta*, a identificación do tipo de desviación do estándar mediante o código L_w_su, a orixe desta desviación, a través do código L_sp, o lema estándar *receita* así como a súa categoría gramatical NCFP000.

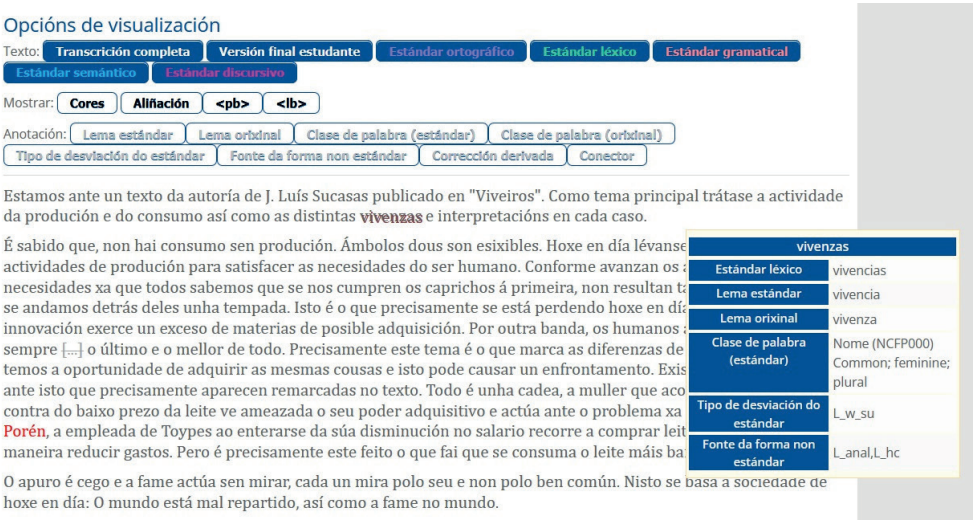
(1) <tok id=»w-137» form=»recetas» lcform=»receitas» olemma=»receta» problem=»L_w_su» psource=»L_sp» lemma=»receita» pos=»NCFP000»>recetas</tok>

No exemplo (2) pode verse a división en <dtok> que se leva a cabo nas contraccións e a asignación aos <dtok> do lema e categoría gramatical:

(2) <tok id=»w-114»>no<dtok lemma=»en» pos=»SP» id=»d-114-1» form=»en»/><dtok lemma=»o» pos=»DA0MS0» id=»d-114-2» form=»o»/></tok>

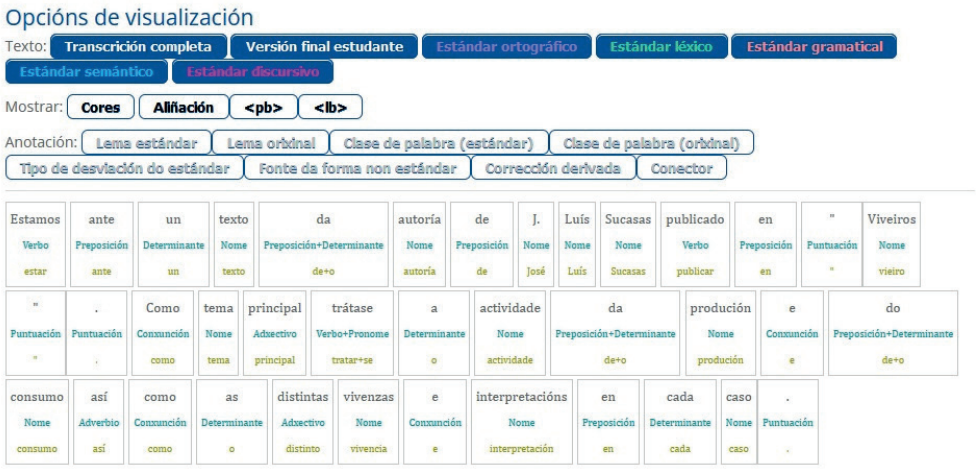
Á hora de visualizar os textos, todas as anotacións lingüísticas asociadas a un *token* son visibles ao colocar o cursor sobre este, tal e como se pode comprobar na Figura 1, en que nos situamos sobre *vivenzas*:

FIGURA 1. Visualización das anotacións lingüísticas dos *tokens* (I)



Ademais, premendo nos botóns das diferentes anotacións (*vid*. Anotación en Figura 1 e 2) estas fanse visibles baixo os *tokens*. Na Figura 2, premeuse en lema estándar e clase de palabras (estándar) no texto anterior (ofrécese só o primeiro parágrafo):

FIGURA 2. Visualización das anotacións lingüísticas dos *tokens* (II)



2.3. Transcrición dos textos

No proceso de transcrición dos textos reproducimos fielmente o manuscrito orixinal e, deste xeito, mantemos as grafías, a puntuación, a acentuación, as abreviaturas e a unión e separación de palabras da/do estudante. Ademais, respectamos a construción de cada texto, reproducindo a súa organización en parágrafos, a súa distribución en páxinas (no caso de que as redaccións se estendan por máis dunha), así como o inicio e o remate de cada liña.

Todas as redaccións son transcritas entre as etiquetas <text> e </text>. Ao longo do texto márcase o comezo de cada liña mediante <lb/> e o remate e inicio dos distintos parágrafos que conforman o texto mediante e . Ademais, tamén se sinalan as páxinas a través da etiqueta <pb/>, que se inclúe aínda que haxa unha soa páxina. Ofrecemos a modo de exemplo un texto transcrito con estas e outras etiquetas, que se comentarán nesta epígrafe:

(3) <text> <pb/><lb/>A gastronomía e a cociña comezaron a ter importancia~~,~~ dende <lb/>sempre, xa que comer foi unha cousa moi importante toda a vida, <lb/>pero co paso do tempo a xente comezou a cambiar de formas <lb/>de pensar, e ~~empezousell<unclear>a</unclear>~~ empezouselle a dar máis importancia <lb/>ao estético, e eu creo que dende eso comezou o cambio na <lb/>gastronomía tamén. Para querer unha cousa primeiro ten que «en<lb/>trar polos ollos», polo que cada paso a cociña comezou a ser máis <lb/>estética, ata tal punto que, a día de hoxe, é máis importante <pb/><lb/>que quede ben no prato a que quite a fame. <lb/>Hai novas técnicas como a deconstrucción, que consiste en se<lb/>parar os elementos dunha comida e cociñalos por separado. <lb/>Tamén se fan concursos de cociña, nos que non só se valora a presen<lb/>tación, tamén

> o sabor e que o prato esté ben cociñado. <lb/>Eu creo que todo eso está ben, innovar e que quede ben no prato <lb/>é algo importante, pero creo que o máis importante é que a co<lb/>mida serve para comer, polo que ten que quitar a fame, e non <lb/>é un xoguete co que se poida provar cousas novas, polo~~<gap quantity=»1» unit=»character»/>~~ que <lb/>penso que a cociña moderna ten o seu punto positivo ~~ata q~~ <lb/>sempre que non ~~<gap quantity=»1» unit=»character»/>~~ sexa para un fin lúdico. </text>

A introdución das etiquetas comentadas permite visualizar o texto na súa forma orixinal, aínda que esta non é a única posibilidade existente, posto que, premendo no botón «Aliñación» das diferentes «Opcións de visualización», a vista do texto pode modificarse. Na Figura 3 vemos o texto de (3) coas liñas dispostas segundo o orixinal e coa indicación do inicio dunha nova páxina mediante unha liña horizontal; na Figura 4, as liñas discorren todo ao longo da pantalla e sen indicación do cambio de páxina e, finalmente, na Figura 5, onde está activada esta última visualización e as opcións <pb> e <lb> de «Mostrar», amósanse barriñas verticais que identifican o inicio e o final de cada liña e un número ([1]) que sinala o remate da páxina e o inicio da seguinte.

Figura 3. Visualización do texto na súa forma orixinal

FIGURA 4. Visualización do texto sen a disposición de liñas orixinal e sen distribución en páxinas

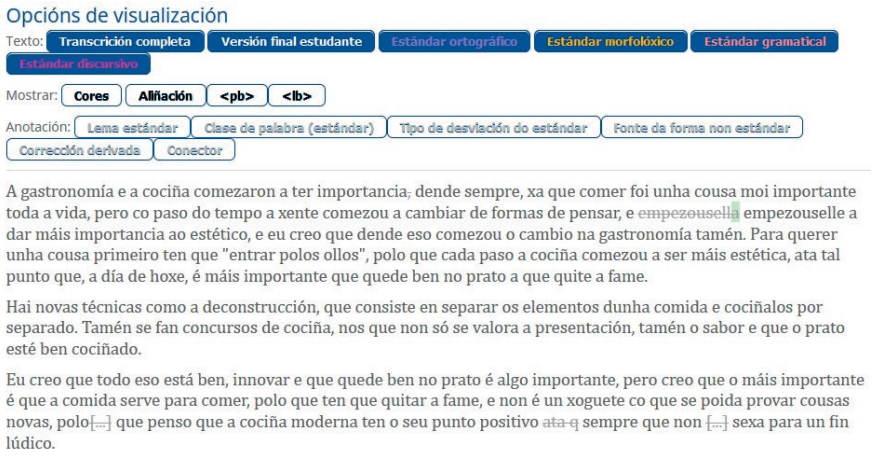
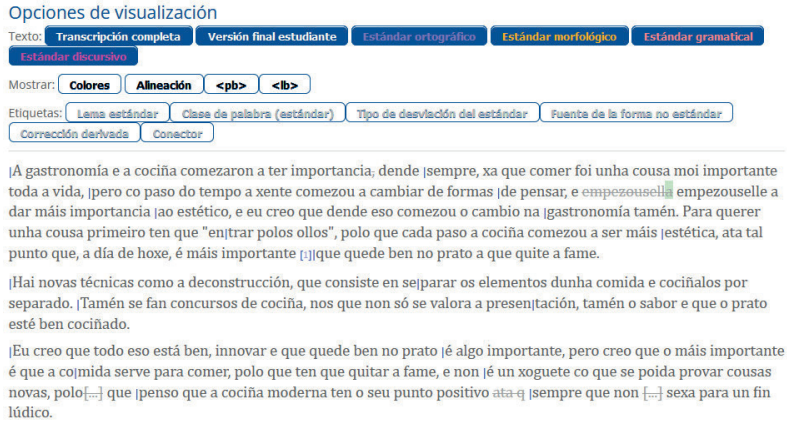


FIGURA 5. Visualización do texto con marcas de inicio e remate de cada liña e páxina

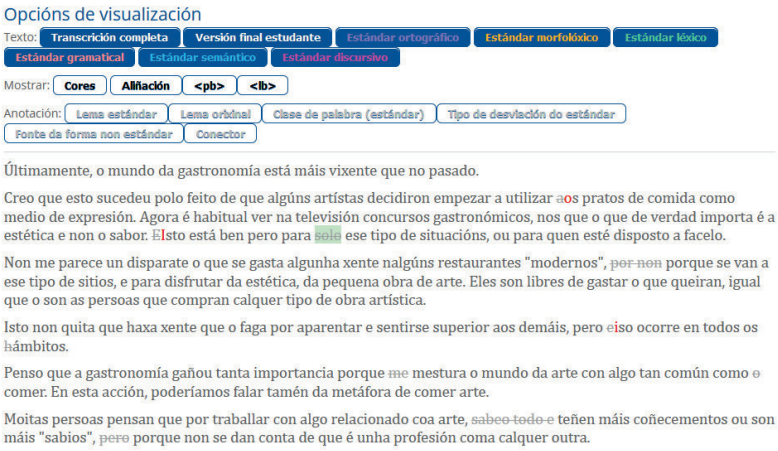


No primeiro tipo de visualización tamén se amosan os cortes de palabra en final de liña, tal e como se pode comprobar en varios *tokens* da Figura 3.

O respecto ao manuscrito de partida tamén se manifesta no rexistro dos engadidos, as riscaduras e as emendas que se aprecian no texto orixinal e que pertencen ao alumno ou alumna, o que, por outro lado, permite rastrexar o seu proceso compositivo. Os fragmentos borrados no texto orixinal mediante calquera sistema (riscado, borrancho etc.) márcanse mediante o elemento <del>. Se o texto borrado é lexible, este incorpórase tras o citado elemento, como na palabra *pero*, riscada neste fragmento: «máis <lb/>»sabios», ~~pero~~ porque non se dan conta de que é unha profesión

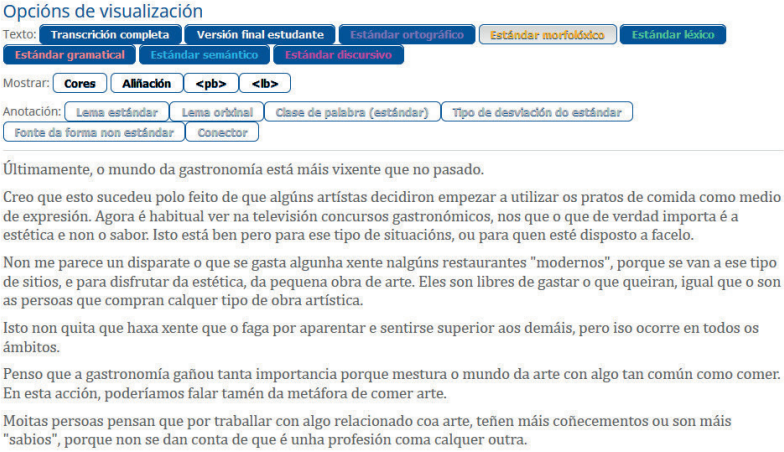
<lb/>coma calquer outra». Escollendo dentro das «Opcións de visualización» o modo «Transcrición completa», os fragmentos eliminados amósanse en gris e riscados, tal e como se pode ver na Figura 6, onde se recollen varias palabras ou secuencias borradas, entre elas a forma *pero* que acabamos de comentar (última liña). Esta visualización responde á establecida na folla de estilo do corpus mediante linguaxe CSS:

FIGURA 6. Visualización das formas riscadas



Premendo, dentro das «Opcións de visualización», en «Versión final estudante», as formas riscadas desaparecen e o texto pode verse na súa forma definitiva, tal e como se pode comprobar na Figura 7:

FIGURA 7. Visualización da versión final da/do estudante nun texto con formas riscadas



Se o texto riscado é ilexible (sexa unha palabra, un carácter, unha secuencia de palabras ou de caracteres), emprégase un elemento <del> e no seu interior un elemento baleiro <gap/>, con indicación, se é posible, do número de caracteres («character») ou palabras («word») riscadas (mediante os atributos @quantity e @unit): «~~<gap quantity=»4» unit=»character»/>~~». Na visualización, esta etiquetaxe dá lugar a ~~[...]~~, tal e como se pode ver na última liña da Figura 3.

Como indicamos, tamén deixamos constancia das formas engadidas despois dunha primeira versión do texto (recoñecibles, por exemplo, porque figuran na interliña, porque se engaden mediante unha nota ou sobre líquido corrector). Etiquétanse sempre mediante un elemento <add>, con independencia do sistema que a/o estudante empregase para introducir a adición: «o resto de produtos se mercan para un <add>mesmo</add> e coa nosa actitude ~~de derroche~~ egoísta». O texto engadido aparece destacado en vermello, tanto na «Transcrición completa», tal e como se pode comprobar na Figura 8, como na «Versión final». Con todo, no caso de que a adición sexa un fragmento dunha palabra, o destacado en vermello só se visualiza na «Transcrición completa».

Figura 8. Visualización da palabra engadida *mesmo*

Finalmente, tamén se etiquetan as palabras cunha lectura dubidosa (estean riscadas ou non) mediante o elemento <unclear>. O usuario visualízaas cun fondo verde azulado, tal e como se pode comprobar na Figura 9: «tendas de roupa locais, que pecharon porque non poden competir, como <unclear>-tamén</unclear> é o culpable da».

Figura 9. Visualización da palabra de lectura dubidosa *tamén*

Para máis detalles sobre o proceso de transcrición dos textos (por exemplo, sobre a combinación de diferentes etiquetas), poden consultarse as instrucións empregadas no proxecto e que figuran na páxina web do corpus (http://ilg.usc.gal/cortegal/downloads/manual-transcricion.pdf).

2.4. Asignación de metadatos

Os metadatos proporcionados en *CORTEGAL* son de tres tipos, atendendo á clasificación proposta por Burnard (2005): metadatos analíticos, descritivos e administrativos. Distribuímolos, empregando etiquetas TEI, en tres elementos: <fileDesc>, <encodingDesc> e <profileDesc>. Baixo o primeiro elemento, incluímos, por unha banda, datos xerais sobre o corpus: o seu nome, a indicación das entidades que o financiaron e información sobre a súa publicación (persoas e entidade responsables do corpus, lugar de publicación, ano, licenza de distribución), así como sobre a orixe dos textos e o seu carácter anónimo. Por outra banda, incluímos tamén información concreta sobre cada texto: o título de cada un deles, seguindo a estrutura ABAU/2016-2017/CD04/xuño/03, onde o número da CD, a convocatoria (xuño ou setembro) e o número do texto dentro de cada CD son variables; as diferentes persoas responsables da transcrición, da revisión da transcrición, da lematización e da anotación lingüística do texto; e finalmente, os seguintes datos cuantitativos sobre este: número de palabras, número de lemas, densidade léxica (que resulta de dividir o número de lemas entre o de palabras), número de enunciados, media de palabras por enunciado, número de palabras do enunciado máis longo, número de palabras do enunciado máis curto, número de parágrafos e media de enunciados por parágrafo.

Para a obtención da maior parte destes últimos datos acudimos á ferramenta *DContado* (Gómez Guinovart e Solla, 2017-2022), á que subimos a versión lexicamente estandarizada dos diferentes textos de *CORTEGAL* para a súa análise cuantitativa. Debe sinalarse que o feito de analizar a versión normalizada no nivel léxico[4], versión necesaria para levar a cabo a lematización e o cómputo de lemas en *DContado*, pode dar lugar a pequenas discrepancias entre o número de palabras do texto orixinal e o número de palabras computadas (por exemplo, cando unha unidade complexa, como *sin embargo*, se estandariza no nivel léxico cunha soa palabra, como *porén*). Por outro lado, en *Dcontado* compútanse como unha soa palabra as contraccións e as combinacións de verbo e clítico, fronte á contaxe individualizada de cada compoñente que realiza *TEITOK*, e ademais, fronte a este, non contabiliza os signos de puntuación, de xeito que o total de palabras computado por esta ferramenta é inferior ao número de *tokens* que mostra a aplicación de *CORTEGAL*. Finalmente, debe terse en conta que a análise en lemas que realiza *DContado* ofrece bastantes problemas no tratamento das contraccións e dos

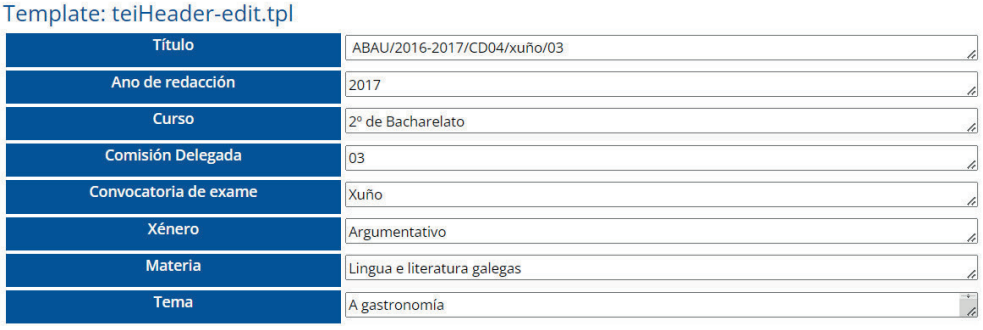
[4] A versión estandarizada no nivel léxico implica estandarización ortográfica, morfolóxica e léxica. *Vid.* §2.5.3.

grupos de verbo e pronome enclítico, de tal xeito que os datos que se ofrecen sobre lemas (e sobre densidade léxica) deben tomarse con cautela e son meramente orientativos.

En segundo lugar, en <encodingDesc> describimos o proceso de transcrición e codificación dos textos e, finalmente, en <profileDesc> informamos doutras características das redaccións, algunhas xerais para todas elas (ano de creación, tipo e fonte dos textos, materia á que pertencen, curso das/dos estudantes) e outras variables, como son a convocatoria, as cualificacións recibidas na pregunta de comentario e no exame de Lingua Galega e Literatura (cualificacións numéricas e cualitativas, neste último caso coas opcións e coas franxas numéricas indicadas en §2.1), o tema dos textos e a CD a que pertencen.

Tras deseñar en *TEITOK* o modelo da cabeceira, a introdución dos metadatos realízase de maneira moi sinxela a través dun formulario cos distintos campos establecidos, nos que se incorpora a información correspondente (pode fixarse un valor fixo para os campos que se desexe). Na Figura 10 ofrécese un fragmento deste formulario:

Figura 10. Fragmento do formulario para a introdución de metadatos



Os datos así introducidos incorpóranse automaticamente na cabeceira de cada arquivo (<teiHeader>), a cal precede o texto (<text>).

No que respecta á visualización dos metadatos, só son visibles para as persoas usuarias do corpus os que se recollen na Figura 11. Por defecto, ademais do título, unicamente se ven a CD, a convocatoria do exame e o tema, pero cun simple «clic» despréganse o resto da información.

Figura 11. Metadatos visibles para as persoas usuarias do corpus

ABAU/2016-2017/CD04/xuño/03

Ano de redacción	2017
Curso	2º de Bacharelato
Comisión Delegada	03
Convocatoria de exame	Xuño
Xénero	Argumentativo
Materia	Lingua e literatura galegas
Tema	A gastronomía
Número de palabras	195
Número de lemas	103
Densidade léxica	53
Número de enunciados	8
Media de palabras por enunciado	24.38
Número de palabras do enunciado máis longo	45
Número de palabras do enunciado máis curto	5
Número de parágrafos	3
Media de enunciados por parágrafo	2.67

No buscador, os metadatos que funcionan como posibles elementos de filtraxe dos textos son o título, o tema, a convocatoria e a CD, así como os datos cuantitativos sobre o texto. As cualificacións só son accesibles ás persoas con permisos de edición. Se se dispón de tales permisos, poden consultarse na cabeceira de cada texto e poden realizarse filtraxes por cualificación a través do buscador.

2.5. Estandarización e codificación das formas non estándar

Neste apartado, tras presentar o marco metodolóxico xeral en que se realiza a estandarización e etiquetaxe das formas non estándar en *CORTEGAL* (§2.5.1) e tras xustificar a realización destas dúas tarefas nos textos do corpus (§2.5.2), ofrecemos unha caracterización do sistema multinivel de estandarización (§2.5.3), mostramos as clasificacións das formas non estándar propostas e os correspondentes sistemas de codificación (§2.5.4), así como algunhas outras anotacións vinculadas coas dúas tarefas indicadas (§2.5.5). Finalmente, presentamos algúns criterios metodolóxicos empregados para minimizar a subxectividade implícita no proceso de estandarización e codificación das formas non estándar e lograr a maior homoxeneidade posible (§2.5.6).

2.5.1. Marco metodolóxico xeral

CORTEGAL sitúase na liña metodolóxica dos xa numerosos corpus de aprendentes elaborados para outras linguas (pode verse unha listaxe dalgúns destes corpus en Centre for English Corpus Linguistics, 2022). Agora ben, a maior parte dos corpus de aprendentes inclúen textos elaborados por estudantes dunha lingua estranxeira ou dunha segunda lingua, e incluso podemos encontrar esta característica como trazo definidor do concepto «corpus de aprendentes»: «[...] learner corpora, which can be defined as electronic collections of natural or near-natural data *produced by foreign or second language (L2) learners* and assembled according to explicit design criteria» (Granger et al., 2015, p. 1; a cursiva é nosa). Porén, de acordo co indicado en §2.1, os textos de *CORTEGAL* foron elaborados por estudantes que teñen o galego como lingua inicial (L1) ou como segunda lingua (L2), entendida esta última como lingua non aprendida no ámbito familiar, pero si en contexto natural e formal desde os primeiros anos de formación académica.

Con todo, cremos que é pertinente considerar *CORTEGAL* como un corpus de aprendentes, posto que o elevado número de formas non estándar atopadas nos textos (*vid.* §3.3) é boa mostra de que o alumnado non domina a destreza da escritura na variedade estándar da lingua galega e esta está aínda en proceso de aprendizaxe. Concordamos, pois, con Abel et al. (2014) na pertinencia de utilizar o termo «aprendentes» tamén para corpus que conteñen textos en L1, como sucede co corpus *KoKo* de textos en alemán (Abel et al., 2012):

> We refer to people as L1 learners when they are still in the process of learning their L1 or related skills of importance such as writing and text production. (...). From a linguistic point of view, the texts written by L1 language learners are likely to have many features of non-standard writing in common with L2/FL learners. However, since some features are specific to either L1 or L2/FL learners, both learner types relate to separate learner varieties. From the perspective of computational processing, L1 and L2/FL learner corpora are fully equivalent since both are compilations of textual data that may deviate from the standard variety (Abel et al., 2014, p. 2414).

Máis en concreto, o noso corpus sitúase no marco metodolóxico coñecido como «Análise informatizada de erros» ou «Análise de erros asistida por ordenador» (Dagneaux et al., 1998), que ten como un dos seus elementos centrais a estandarización e codificación das formas ou secuencias non estándar, aspecto que trataremos nas seguintes subepígrafes de §2.5.

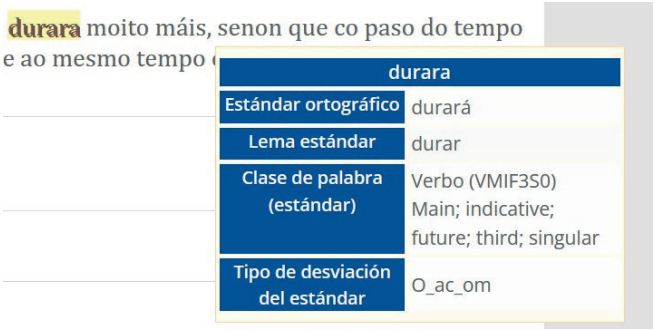
2.5.2. Dous procesos: estandarización e codificación

Tal e como indican Díaz-Negrillo e Fernández-Domínguez (2006, p. 86), nos corpus de aprendentes con análise informatizada de erros é habitual combinar dous procesos complementarios: a introdución de etiquetas identificadoras de tipos de formas desviantes[5] e a normalización ou asignación das formas estándar correspondentes, denominadas frecuentemente *target hypothesis* (Reznicek et al., 2013, p. 104, Stemle et al., 2019, p. 428). Malia certas reticencias cara á normalización pola imposibilidade por parte do/da investigador(a) de coñecer as intencións da persoa que escribe os textos (Alonso-Ramos, 2016, p. 8), como sinalan Lüdeling e Hirschmann (2015, p. 141), «an error-annotated corpus which does not provide target hypotheses hides an essencial step of the analysis», pois a codificación de cada forma non estándar está condicionada e xustificada, cando menos en parte, pola forma estándar asignada en cada caso. De acordo con este criterio, en *CORTEGAL* levamos a cabo tanto a etiquetaxe das formas non estándar como, na maior parte dos casos[6], a súa normalización ou estandarización, tal e como se pode comprobar na Figura 12, en que se corrixe *durara* por *durará* (en «Estándar ortográfico») e ao mesmo tempo se asigna un código identificador do tipo de desviación do estándar que corresponde, unha omisión do acento gráfico (O_ac_om [Ortography_accent_omission]). En *CORTEGAL*, os procesos de estandarización e codificación realízanse simultaneamente e non en fases sucesivas.

[5] Aínda que en certos casos, como por exemplo nas desviacións ortográficas, o emprego da palabra *erro* non supón ningún problema, cando falamos do conxunto de formas anotadas en *CORTEGAL* preferimos empregar formas non estándar ou formas diverxentes/desviantes do estándar ou da norma, en vez de usar o termo *erros*, palabra que, polas súas connotacións, resulta pouco apropiada para a etiquetaxe, por exemplo, dos dialectalismos. Incluímos ademais etiquetas para identificar voces ou construcións que se desvían do estándar académico polo seu rexistro, pero que son perfectamente correctas en niveis de lingua coloquiais e tamén neste caso o termo *erro* parece pouco afortunado. Tal e como sinalan Alonso-Ramos (2016, p. 8) ou Stemle et al. (2019, p. 15) xa noutros casos se ten evitado o emprego desta palabra, substituída, por exemplo, por *uso non convencional*, *desviación da norma* etc.

[6] Unha das etiquetas que utilizamos informa da ininteligibilidade de todo un enunciado (*vid.* §2.5.4). Nestes casos, non ofrecemos forma normalizada. Tampouco o facemos naqueles casos en que é evidente que se está utilizando unha palabra cun significado que non lle corresponde, pero no que nos resulta imposible determinar que contido se quere transmitir.

FIGURA 12. Anotación e normalización da forma *durara* nun texto de *CORTEGAL*



Con respecto ao obxecto da estandarización e da codificación, normalízanse e etiquétanse aquelas formas, usos ou omisións que non corresponden ao código normativo galego, que son inadecuados para un rexistro formal e académico como o que corresponde ao das probas ABAU ou que non respectan as convencións establecidas para textos deste tipo, atendendo aos niveis ortográfico, morfolóxico, semántico, léxico, gramatical (sintáctico) e discursivo. En §2.5.3 pode atoparse unha descrición máis detallada dos aspectos que son estandarizados e etiquetados no corpus.

2.5.3. Un sistema multinivel

Tal e como sinalan Stemle et al. (2019, p. 441), «Although very few learner corpus projects have managed to reuse each other's error taxonomies so far, several projects have tried to build on previous work». *CORTEGAL* identifícase perfectamente con esta afirmación, posto que, se ben o seu sistema de estandarización e codificación das formas non estándar toma como punto de partida sistemas empregados noutros corpus, utiliza un método propio, construído non só sobre a base dos sistemas anteriores, senón tamén a partir da análise dos problemas detectados nos textos de *CORTEGAL* e tendo en conta as especificidades e posibilidades de *TEITOK*.

Unha das principais diferenzas entre os distintos corpus ten que ver cos niveis cubertos polas taxonomías empregadas na anotación das formas desviantes, entre os cales o ortográfico, o gramatical e o léxico adoitan estar presentes:

> the linguistic levels more commonly covered by the error taxonomies are spelling, grammar and lexis. On the other hand, classifications of phonetic, pragmatic or discoursal errors do not seem to be always present in error tagging systems, and when present, their error categories are rather limited (Díaz-Negrillo e Fernández Domínguez, 2006, p. 89).

No caso particular de *CORTEGAL*, o número de niveis establecidos é de seis, como xa adiantamos e como comentaremos con máis detalle nesta epígrafe. Ademais, o sistema de estandarización utilizado en *CORTEGAL* é, como noutros corpus (por exemplo, *COPLE2*, Mendes et al., 2016; *KoKo*, Abel et al., 2016; ou *Falko*, Reznicek et al., 2013) un sistema multinivel, que permite a asignación de formas normalizadas en diferentes dimensións lingüísticas (Díez-Bedmar, 2021, p. 97).

Algúns corpus multinivel, como o *CzeSL-man* (Rosen, 2015), *Falko* (Reznicek, 2012) ou *MERLIN* (MERLIN project, 2014) permiten a asignación de *target hypotheses* tan só en dous niveis. En trazos xerais, no primeiro nivel lévanse a cabo estandarizacións de carácter esencialmente formal, como as ortográficas e as gramaticais cando afectan a elementos illados, mentres que no segundo nivel se realizan outras modificacións máis complexas que teñen en conta o contexto, entre elas as semánticas e as estilísticas.

Noutros corpus multicapa, como o *CroLTeC* (Mikelić Preradović, 2020) ou o *COPLE2* (del Río e Mendes, 2018) escóllense os tres niveis «clásicos» indicados por Díaz-Negrillo e Fernández Domínguez na cita anterior: ortográfico, gramático e léxico. Outros engaden un cuarto nivel relativo aos aspectos discursivos, como o corpus *KoKo*, cunha capa para a estandarización da puntuación.

Outro aspecto que pode diverxer entre os distintos corpus ten que ver coa distribución dos tipos de desviación entre os distintos niveis: por citar un exemplo, o corpus *KoKo*, tal e como indicamos, asigna un nivel específico para os problemas de puntuación, mentres que outros corpus, por exemplo o *COPLE2*, os integra dentro da ortografía (Amaro et al., 2020, p. 14).

De acordo co indicado, non existe unha proposta única de niveis lingüísticos nos corpus multicapa. No caso de *CORTEGAL*, tal e como indicamos, decidimos establecer seis niveis distintos, ordenados do seguinte xeito: ortográfico, morfolóxico, léxico, semántico, gramatical (sintáctico) e discursivo. O elevado número de niveis lingüísticos establecidos xustifícase pola posibilidade de que unha mesma forma reciba formas normalizadas distintas e incompatibles entre si en varios deses niveis, tal e como amosaremos máis abaixo nesta mesma epígrafe.

No nivel ortográfico, anotamos problemas relativos ao emprego de signos diacríticos (acentuación e diérese), de maiúsculas e minúsculas, á escritura conxunta ou separada das palabras, á confusión, adición ou omisión de letras ou dígrafos, á escritura de estranxeirismos e abreviaturas ou siglas e á representación de contraccións e asimilacións. Sobre os problemas de acentuación, véxase o capítulo de López-Sández e Lorenzo-Herrera neste volume.

No que respecta ao nivel morfolóxico atendemos exclusivamente a problemas de flexión, sexa de verbos, de substantivos, de adxectivos ou de palabras gramaticais. Deste xeito, estandarízanse e codifícanse neste nivel cuestións relativas ao xénero, ao número, á conxugación verbal e aos morfemas avaliativos.

No nivel léxico, estandarizamos e etiquetamos os casos de emprego dunha unidade léxica con diferenzas de xénero ou de acentuación de intensidade con respecto ao galego estándar, en ambos os casos xeralmente por influencia do cognado español (por exemplo, *leite*, con 15 exemplos en feminino por influencia do español *leche*, que ten este xénero; ou *élite*, con 7 exemplos, por influencia do español *élite*, fronte ao galego estándar *elite*). Pero, sobre todo, inclúense aquí outras desviacións léxicas en que as diferenzas co estándar van máis alá de modificacións no xénero ou na acentuación, como o uso de *receta* en vez de *receita*, de *abandoar* no canto de *abandonar*, de *actitud* por *actitude*, de *todavía* por *aínda* etc. Sobre as formas non estándar do nivel léxico *vid.* o capítulo de Álvarez de la Granja nesta obra.

No nivel gramatical atendemos a cuestións relacionadas coa sintaxe da frase e da oración: problemas de concordancia, omisión e adición de preposicións, problemas no emprego de pronomes, determinantes e conxuncións, selección gramaticalmente inadecuada de tempo, modo ou voz verbal, construción non estándar dun esquema verbal, escolla dunha clase de palabra inadecuada ou, en xeral, presenza dunha estrutura sintáctica agramatical. Véxanse o capítulo de Fernández Salgado sobre a colocación dos pronomes átonos e o de Cidrás sobre cuestións de concordancia, neste volume.

No nivel semántico anotamos os usos ou significados dunha palabra que non son estándar ou adecuados ao contexto ou ao sentido que plausiblemente se quere transmitir (por exemplo, o emprego de *causas* por *consecuencias* en «provocar causas nefastas»), así como a omisión de palabras necesarias para o sentido do texto ou a adición de voces ou expresións innecesarias, normalmente pola súa redundancia.

Finalmente, no nivel discursivo rexistramos os problemas relativos á puntuación, incluíndo a división do texto en parágrafos, e ao emprego de conectores e de partículas referenciais, así como as elipses inadecuadas, a orde discursivamente inapropiada, aínda que gramaticalmente aceptable, os problemas relativos ao rexistro das palabras ou construcións e, finalmente, os enunciados demasiado complexos e os inintelixibles.

Con respecto aos problemas de puntuación, que nalgúns corpus, como *COPLE2* ou *MERLIN*, se consideran dentro do nivel ortográfico, optamos

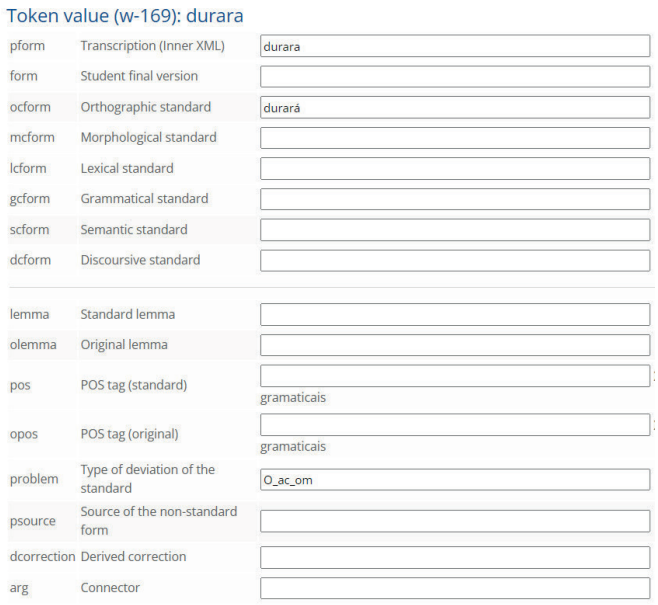
por incluílos no nivel discursivo na medida en que os signos de puntuación son un elemento fundamental para a organización do texto. Tal e como sinala Figueras (1999), «cabe entender la puntuación como un mecanismo más de organización de la información del texto; su función es delimitar y articular las diversas unidades textuales de procesamiento». Do mesmo xeito, Roselló Verdeguer (2015), quen considera os signos de puntuación como elementos de cohesión discursiva, incide na necesidade de separar ortografía e puntuación:

> Los signos de puntuación han recibido poca atención por parte de la lingüística, que los ha considerado un aspecto de la ortografía a caballo entre la oralidad y la escritura. Aunque es cierto que la primitiva función de los signos era señalar los lugares donde el lector debía realizar las pausas, hoy se considera la puntuación como un elemento de la escritura, y, por tanto, un mecanismo para articular el contenido del texto y dotarlo de una estructura.

Tal e como indicamos, integramos, pois, os problemas de puntuación cos que afectan a outros mecanismos de cohesión, que, canda a puntuación, contribúen a dotar o texto de coherencia, como a conexión, as elipses e a referencia. Somos conscientes, en calquera caso de que ás veces os problemas de puntuación responden a unha mera cuestión ortográfica (por exemplo, a omisión dun punto ao final do texto) ou de que outras veces os signos de puntuación teñen unha dimensión máis gramatical ca discursiva, ao contribuír, por exemplo, á organización da frase. En calquera caso, e para non complicar en exceso o proceso de anotación, todos os problemas relativos á puntuación (agás a omisión de punto en abreviaturas, que se anota no nivel ortográfico coa etiqueta O_ab_su), son codificados na capa discursiva.

Tras establecer o número de capas de estandarización que interesaba no deseño xeral do corpus, a normalización das formas levouse a cabo, na maior parte dos casos (*vid.* §2.5.4 para excepcións), nun formulario de edición de *tokens* mediante a asignación, no nivel correspondente, da forma estandarizada. Así, por exemplo, retomando o exemplo de *durara* que amosamos na Figura 12, e dado que nos atopamos cun problema de acentuación gráfica, debemos escribir a forma correcta na capa de estandarización ortográfica, tal e como se ve na Figura 13. Ademais de levar a cabo a corrección, asignamos un código que identifica o tipo de desviación (*vid.* §2.5.4). A forma estandarizada, e o código correspondente, convértense en atributos do *token*, tal e como indicamos en §2.2.

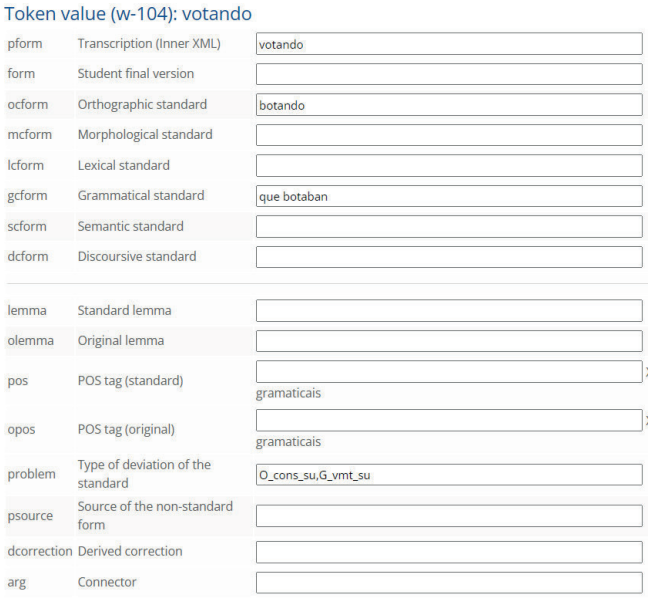
FIGURA 13. Normalización e codificación en *TEITOK*



Nótese que, se non se indica nada nas capas de estandarización, estas van herdar sempre a última forma normalizada e, de non haber normalización de ningún tipo, a forma escrita no orixinal. Deste modo, neste caso, os niveis morfolóxico, léxico, gramatical, semántico e discursivo herdan a forma corrixida ortograficamente *durará*.

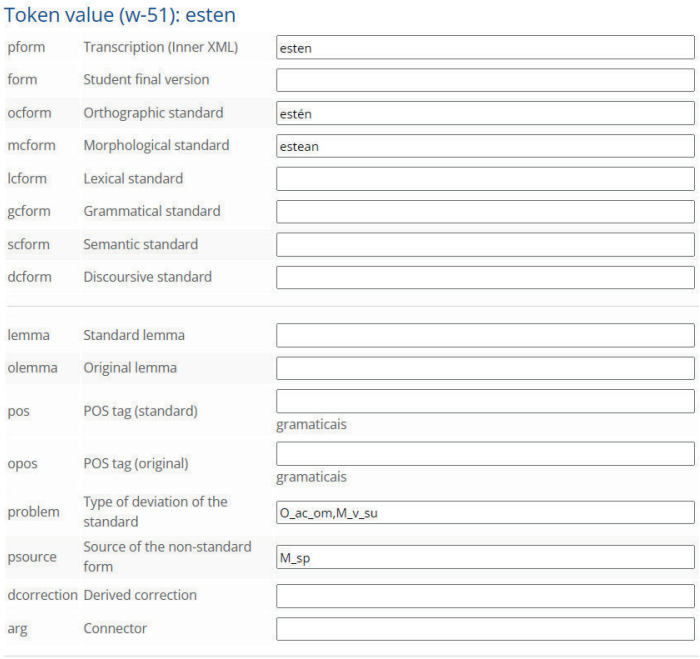
Tal e como sinalamos, a utilización dun sistema multinivel permite a asignación de varias formas normalizadas en capas distintas. Nestes casos, sempre que é posible e pertinente, a estandarización realizada nun nivel mantense nos seguintes. Así, por exemplo, no fragmento «o caldo galego era o medio de subsistencia das clases mais desfavorecidas, votando os restos das comidas no puchero», substituímos no nivel ortográfico *votando* por *botando* e no nivel gramatical, en que se muda o xerundio por unha oración de relativo, conservamos a corrección ortográfica («que botaban» e non «que votaban»), tal e como se ve na Figura 14.

FIGURA 14. Dobre estandarización e conservación das modificacións en *TEITOK*



Agora ben, nalgúns casos tal conservación non é posible na medida en que as diferentes formas normalizadas non son compatibles entre si. Así, por exemplo, tal e como se ve na Figura 15, *esten*, terceira persoa de plural do presente do subxuntivo do verbo *estar*, corríxese no nivel ortográfico mediante *estén*, dado que é unha palabra aguda rematada en <-n> que debería levar acento gráfico, pero estandarízase de novo, desta volta mediante *estean*, na capa morfolóxica, dado que *estén* é unha forma flexionada non estándar, producida por influencia do español. Asignamos ademais os códigos correspondentes que identifican o tipo de desviación e a orixe da diverxencia morfolóxica, tal e como se explicará en §2.5.4.

FIGURA 15. Dobre estandarización sen conservación das modificacións en *TEITOK*



Nótese que a forma *estean* non mantén, por incompatible, a corrección ortográfica relativa á acentuación e que se traballásemos nun sistema cun só nivel de *target hypotheses*, estandarizando directamente *esten* por *estean*, sen ofrecer como paso intermedio a forma normalizada *estén*, estariamos agochando un paso fundamental da análise, aquel que xustamente explica a asignación do código de omisión de acento (O_ac_om). Precisamente, e tal e como indicamos antes, a proposta de seis niveis lingüísticos xustifícase polo feito de que cada un deles pode requirir estandarizacións diferenciadas e incompatibles, o que nos leva a distinguir, por exemplo, entre o nivel morfolóxico (flexión) e o gramatical (sintaxe da frase e a oración), niveis que nalgúns corpus, como *KoKo* e *MERLIN*, se unifican. No exemplo que se ofrece a continuación, encontramos a acumulación nunha mesma forma de problemas dos dous tipos, aos que corresponden formas normalizadas diferentes en cada nivel:

(4) Tamén hai que ter en conta que no noso caso, en España, dende fai uns poucos anos na televisión *esteñan* a desenrolar programas exclusivamente de concursos de cociña

En (4), no nivel morfolóxico débese substituír a variante dialectal *esteñan* do presente de subxuntivo pola forma estándar *estean*, mentres que no nivel gramatical existe un problema na selección do modo verbal, de modo que é necesario empregar o indicativo: *están*. A substitución directa por *están* agocharía a forma normalizada *estean* que xustifica a asignación da etiqueta morfolóxica M_v_su.

Polo mesmo motivo, e fronte á práctica relativamente habitual nos corpus de aprendentes (por exemplo *COPLE2* ou *MERLIN*), diferenciamos entre o nivel léxico e o semántico, posto que en varios casos é preciso atribuír formas normalizadas diferentes e incompatibles para cada nivel, tal e como ocorre en (5):

(5) Os costes de producción son moi *amplios*

No fragmento anterior, o castelanismo *amplios* debe estandarizarse no nivel léxico por *amplos*, pero no nivel semántico é preciso levar a cabo outra modificación, posto que *amplos/amplios* non é un adxectivo apropiado ao contexto (fronte a, por exemplo, *elevados*, que empregamos como forma normalizada).

Os problemas de rexistro, codificados nalgúns corpus no nivel léxico (por exemplo no corpus *KoKo*), son tamén separados deste por idénticas razóns. Así, por exemplo, en (6)

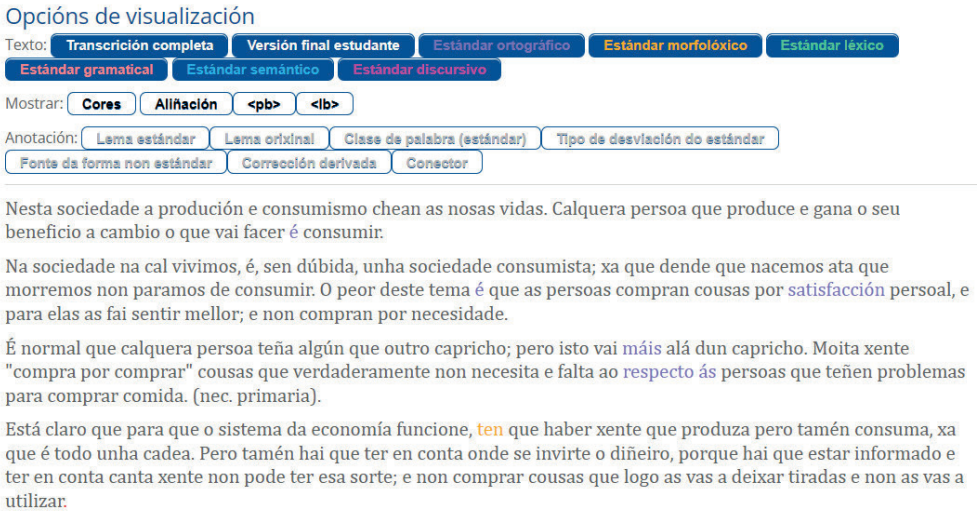
(6) Teñen unha pinta *asquerosa*

cómpre estandarizar o castelanismo *asquerosa* por *noxenta* no nivel léxico, pero consideramos necesaria unha segunda modificación no nivel discursivo, na medida en que nin *noxenta* nin *asquerosa* parecen formas adecuadas ao rexistro formal dos textos de *CORTEGAL*, polo cal propoñemos *desagradable*.

TEITOK, por outro lado, permite visualizar os textos nas diferentes capas de estandarización, tendo en conta que en cada capa se ven as modificacións propias dese nivel e as dos anteriores. Así pois, na última capa, a discursiva, obtemos o texto estandarizado en todas as dimensións lingüísticas. Ademais, as formas normalizadas de cada nivel teñen asignadas en *CORTEGAL* cores distintas para poder distinguir claramente as unhas das outras. Así, por exemplo, na Figura 16 visualizamos un dos textos tras premer na capa de estandarización morfolóxica («Estándar morfolóxico» en «Opcións de visualización»), de tal xeito que vemos as modificacións realizadas no nivel

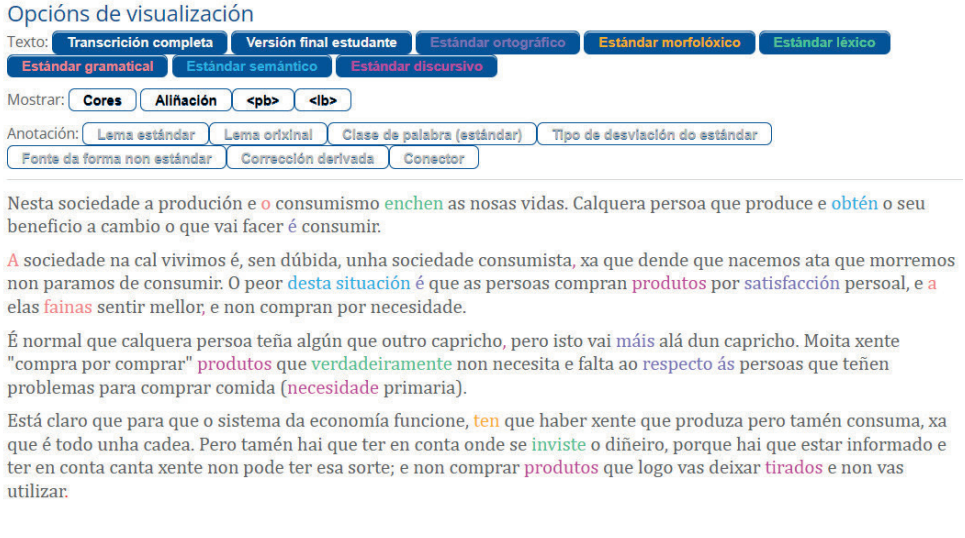
morfolóxico (un único cambio, en laranxa) e no nivel anterior, o ortográfico (en lila):

FIGURA 16. Visualización dun texto na capa de estandarización morfolóxica



En cambio, na Figura 17, o mesmo texto visualízase na capa discursiva (tras premer en «Estándar discursivo»), de tal xeito que ademais das formas normalizadas anteriores, vemos as modificacións do nivel léxico (en verde), do gramatical (en salmón), do semántico (en azul) e do discursivo (en fucsia).

FIGURA 17. Visualización dun texto na capa de estandarización discursiva



Do exposto ata o de agora dedúcese que a orde establecida para as capas de estandarización non é en absoluto irrelevante, posto que as modificacións se realizan e se visualizan de acordo con esa orde, comezando polas mudanzas formais a nivel de palabra (ortografía, morfoloxía e léxico: primeiro as meramente gráficas, despois as que afectan á flexión e finalmente as que corresponden á selección da unidade léxica), seguindo polas formais a nivel da frase ou da oración (gramática), continuando coas relacionadas co significado (semántica) e rematando con aquelas que están vinculadas coa construción do texto e coa adecuación ao contexto (discurso).

2.5.4. Os códigos de anotación das formas non estándar

Tal e como xa comentamos e como queda patente en numerosos traballos (véxase, por exemplo, Díaz-Negrillo e Fernández Domínguez, 2006; Stemle et al., 2019 ou Díaz-Bedmar, 2021), a taxonomía para a clasificación das formas desviantes do estándar que atopamos naqueles corpus de aprendentes que ofrecen este tipo de anotación pode variar considerablemente duns proxectos a outros. Estas diferenzas encóntranse nos criterios de clasificación empregados, na granularidade da taxonomía ou nos niveis ou aspectos lingüísticos considerados.

As clasificacións e estudos realizados en diferentes traballos antes do nacemento da análise informatizada de erros (entre eles Corder, 1974; Dulay et al., 1982 ou Ellis, 1994) serviron como base para as taxonomías aplicadas nos corpus informatizados. Deste xeito, é frecuente atopar clasificadas as formas non estándar atendendo ao nivel lingüístico ou ao aspecto concreto ao que afecta a desviación e tamén tendo en conta as modificacións que se producen con respecto ao estándar (omisións, adicións, cambios de orde...). Menos frecuentemente, tamén se poden encontrar clasificacións que consideran as causas da desviación (erros interlingüísticos, intralingüísticos...).

No caso de *CORTEGAL*, a clasificación das formas desviantes do estándar realízase atendendo aos tres aspectos mencionados, en dúas taxonomías separadas. A primeira tenta describir a forma non estándar combinando información sobre o nivel e o aspecto afectado con información sobre as modificacións que se producen con respecto ao estándar e a segunda, só empregada en certos casos, dá conta da orixe ou causa da desviación.

Na primeira das taxonomías, empregamos códigos conformados por tres elementos separados por barra baixa, como, por exemplo, O_ac_om. O primeiro elemento, sempre en maiúscula, indica o nivel lingüístico ao que afecta a desviación: O: Ortografía, M: Morfoloxía, L: Léxico, G: Gramática (sintaxe), S:

Semántica e D: Discurso. O segundo elemento, o que ofrece maior variación, concreta o aspecto afectado dentro do nivel lingüístico xeral (*ac*: acentuación, *gen*[7]: xénero, *ref*: elementos con valor referencial, *prep*: preposicións etc.). Como se pode observar, tales elementos poden ser clases de palabras, trazos lingüísticos, categorías funcionais etc. Cando queremos indicar unha unidade léxica en xeral usamos «w» («word»), como en L_w_su (substitución dunha unidade léxica estándar por unha non estándar). Finalmente, o terceiro elemento explica a diferenza existente entre a forma estándar e a forma proporcionada pola/o estudante, sempre desde esta última perspectiva, ou o trazo que fai que a forma ou secuencia escrita non sexa estándar (*ad*: adición, *om*: omisión, *su*: substitución, *wp*: posición incorrecta ou inadecuada, *tr*: transposición, *un*: unión, *spl*: división, *com*: excesiva complexidade, *unin*: inintelixibilidade).

O número total de códigos é de 80, distribuídos do xeito que se indica na Táboa 3, onde se pode observar que os niveis ortográfico, gramatical e discursivo son os que contan cun maior número de etiquetas. A listaxe completa de códigos, con explicacións, exemplos e aclaracións sobre o seu uso, pode atoparse no manual empregado para a etiquetaxe e estandarización e que se encontra dispoñible na páxina do proxecto (http://ilg.usc.gal/cortegal/downloads/manual-anotacion.pdf).

Táboa 3. Distribución dos códigos descritivos das formas non estándar por nivel lingüístico

Nivel	**Número de códigos**	**Porcentaxe**
Ortográfico	25	31,25%
Gramatical	24	30%
Discursivo	18	22,5%
Morfolóxico	6	7,5%
Semántico	4	5%
Léxico	3	3,75%

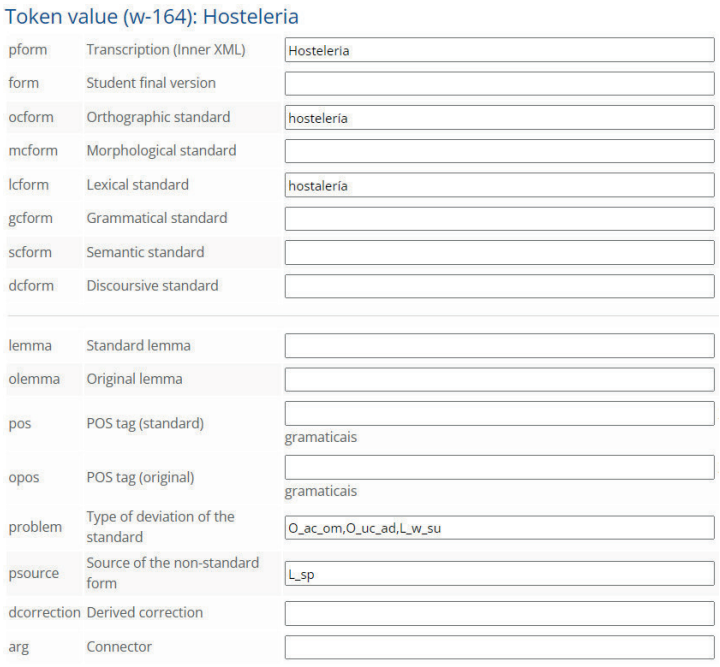
Somos conscientes de que se trata dun número elevado de códigos e de que este feito dificulta a anotación e pode afectar á súa coherencia (*vid*. Díaz-Bedmar, 2021, p. 92). Con todo, tales códigos están construídos mediante a combinación dun número máis reducido de elementos (6 do primeiro nivel, 37 do segundo e 9 do terceiro), o que, no sentido contrario, facilita a anotación.

[7] As abreviacións están creadas sobre as formas inglesas, pensando na posible replicabilidade dos códigos en corpus doutras linguas.

Ademais, dado que o corpus está pensado non só para a investigación, senón tamén para o seu emprego directo nas aulas, unha etiquetaxe demasiado xeral podería ser pouco útil neste contexto. Optamos, daquela, por unha maior granularidade que permita unha clasificación máis desenvolvida das formas desviantes e, daquela, que facilite a localización por parte do profesorado e do alumnado de exemplos de desviacións relativamente concretas.

Como xa indicamos en §2.2, a maior parte das anotacións que estamos comentando son realizadas a nivel de *token*, mediante o mesmo formulario empregado para a anotación das formas normalizadas que presentamos en §2.5.3. Unha mesma forma que acumule varias desviacións recibirá varios códigos, pertencentes ao mesmo ou a distintos niveis, tal e como se ve na Figura 18, en que, na palabra *Hostelería*, no fragmento «a Hosteleria gústalle a xente» se acumulan dúas anotacións no nivel ortográfico (omisión de acento gráfico [O_ac_om] e adición de maiúscula [O_uc_ad]) e unha no nivel léxico (emprego de *hostelería* por *hostalería*) [L_w_su].

Figura 18. Asignación de varios códigos de forma non estándar a unha mesma palabra



Nótese, porén, que esta multiplicidade de códigos está asociada sempre á acumulación de diferentes desviacións, posto que para cada unha delas

ofrecemos un único código e nunca varias interpretacións (sobre esta cuestión, *vid.* §2.5.6).

Un total de 7 dos 80 códigos arriba mencionados non se anotan a nivel de *token*, senón en arquivos separados ligados ao texto, empregando un sistema *stand-off*. Isto sucede con aqueles códigos que afectan ou poden afectar a unha secuencia de *tokens*, que son os que identifican as reformulacións e topicalizacións que dan lugar a estruturas gramaticais inaceptables (G_str_ref-top_su), as adicións de palabras que orixinan tamén unha estrutura agramatical (G_w_ad), as estruturas agramaticais que non encaixan noutros códigos (G_str_su), o cambio de orde discursivamente inadecuado (D_w_wp), os enunciados excesivamente complexos (D_ut_com) e os ininterlixibles (D_ut_unin), así como a adición de elementos innecesarios, normalmente por seren semanticamente redundantes (S_w_ad). A razón desta diferente anotación explícase na páxina de *TEITOK* e radica nas dificultades de asignar un código a toda a secuencia se os seus compoñentes xa posúen outras anotacións:

> Whenever more information needs to be encoded in the corpus that spans multiple words, and which can with [overlap] other types of information, there is not way to encode such information inside the XML file. The only way to encode such information is what is called stand-off: stored in a separate file which is linked to the TEI document[8].

Debe terse en conta que os códigos indicados se anotan sempre mediante un sistema *stand-off*, independentemente de que nalgúns casos poidan afectar unicamente a unha palabra. Así, por exemplo, a etiqueta S_w_ad pode atinxir a unha secuencia de palabras, como en (7), en que sobra o elemento parentético que figura ao final, pois resulta redundante, ou a unha única palabra, como en (8), onde sobra *quizá* (os destacados son nosos):

(7) A gastronomía e a cociña, nos últimos anos acadaron moita popularidade, aínda que hai que recoñecer, que polo menos en Galicia sempre foi e continúa a ser dos trazos que máis caracterizan o noso patrimonio e cultura *(no ámbito gastronómico)*
(8) É probable que *quizá* a estética dos platos evolucione

Só desta maneira aseguramos que o conxunto de exemplos codificado coa mesma etiqueta (neste caso con S_w_ad) se recupere conxuntamente, posto que os códigos asignados mediante este sistema de anotación deben ser buscados nun apartado específico do buscador de *TEITOK* («Anotacións multipalabra»).

[8] http://www.teitok.org/index.php?action=help&id=standoff

Por outro lado, tal e como se ve na Figura 20, as estandarizacións realizadas mediante un sistema *stand-off* son visibles premendo na ligazón «Anotación multipalabra» que aparece naqueles textos que contan con este tipo de anotacións (destacado cun rectángulo vermello na Figura 19):

FIGURA 19. Ligazón para consultar as anotacións multipalabra

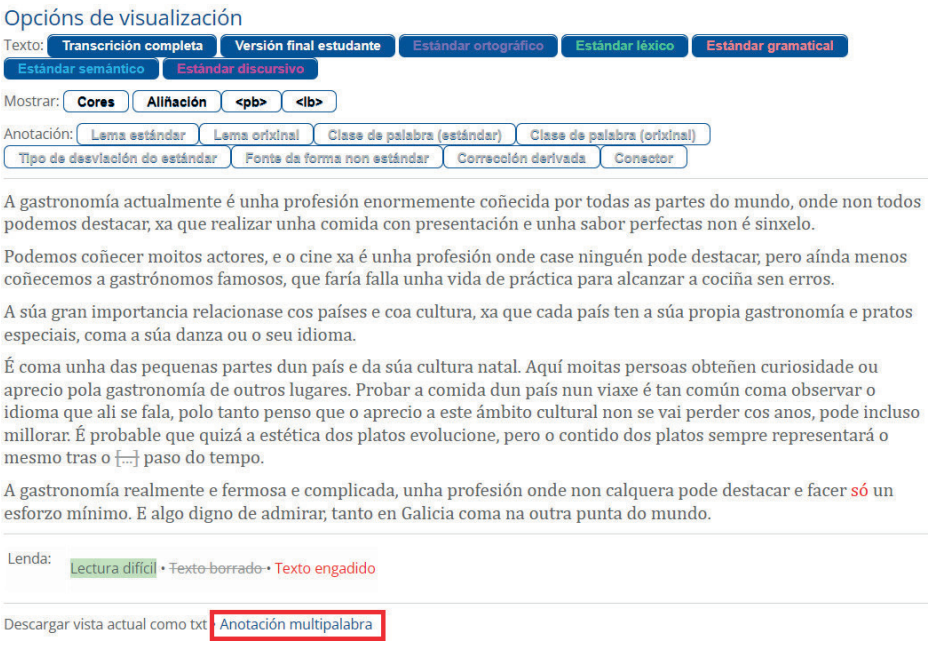
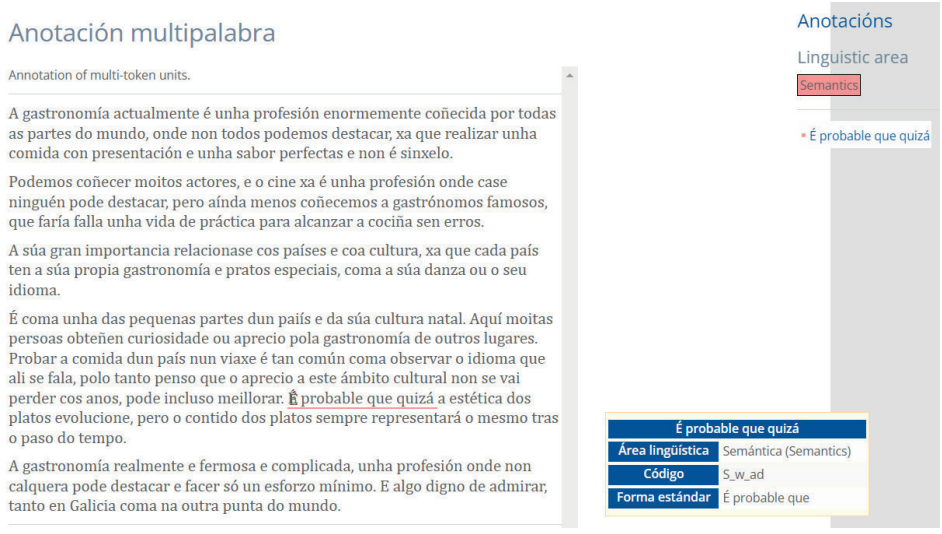


FIGURA 20. Pantalla que se visualiza tras premer na ligazón Anotación multipalabra



Porén, estas estandarizacións non se actualizan nas distintas capas de visualización mencionadas en §2.5.3. Por tal motivo, e co obxectivo de que estas capas ofrezan unha versión estándar completa do texto, ademais de informar da forma modificada en *stand-off*, levamos a cabo a nivel de *token* a mudanza individual de cada unidade afectada (sen asignar códigos, evidentemente, que corresponden a toda a secuencia e só son asignables en *stand-off*), ata obter a versión final normalizada. Por este motivo, algunhas palabras que aparecen estandarizadas nos textos non están asociadas, a nivel de *token*, a ningún tipo de código.

A anotación en *stand-off*, limitada aos códigos indicados, debe diferenciarse da anotación de unidades complexas creadas en *TEITOK* mediante a función *merge* que ofrece a plataforma, e que empregamos cando é preciso estandarizar e codificar en bloque unha expresión complexa (pode ser unha expresión lexicalizada ou calquera outra unidade que realice unha función unitariamente e deba ser substituída en conxunto). Así, por exemplo, unificamos os compoñentes da locución substantiva *perrito quente*, que se estandariza e etiqueta conxuntamente con *bocadillo de salchicha* e L_w_su. En calquera caso, restrinximos o máximo posible o emprego desta función, que ofrece algunhas limitacións e certas dificultades na recuperación a través do buscador.

A anotación que comentamos ata o de agora serve para describir o tipo de desviación con respecto ao estándar, pero, tal e como indicamos ao comezo deste apartado, existe unha segunda taxonomía, moito máis simple, que ten como función ofrecer posibles explicacións sobre a diverxencia da norma. Porén, este tipo de anotación é moito máis subxectiva, tal e como sinalan Lüdeling et al. (2005), e en moitas ocasións resulta realmente difícil determinar a orixe da forma desviante. Por tal motivo, a aplicación deste código segue criterios diferentes dos empregados no uso do código descritivo. Por un lado, non todas as formas desviantes reciben unha etiqueta explicativa e, por outro lado, en ocasións ofrecemos máis dun código deste tipo para explicar unha mesma desviación. As etiquetas empregadas, os niveis en que se aplica, o seu valor e algúns exemplos figuran na Táboa 4:

Táboa 4. Códigos empregados para explicar a orixe da forma diverxente

Orixe	**Valor**	**Códigos**	**Exemplos**
anal	Formas que responden a un proceso de analoxía con outras palabras da mesma lingua	O_anal M_anal L_anal	O_anal: *hippie* (sobre o plural *hippies*)... M_anal: *puidendo* (sobre o tema do pretérito *puid-*)... L_anal: *nutrinte* (a partir de voces como *ferinte* ou *seguinte*)...
for	Palabras cuxa desviación se pode explicar pola súa orixe estranxeira e calcos de voces estranxeiras. Exclúense sempre as formas con orixe no español. No ámbito ortográfico, palabras cunha ortografía incorrecta. No ámbito morfolóxico, palabras con flexión non estándar. No ámbito léxico, palabras que non presentan a adaptación proposta no estándar. No ámbito semántico, calcos semánticos.	O_for M_for L_for S_for	O_for: *graffitti* (por *graffiti*)... M_for: *chef* (por *chefs*) L_for: *tablet* (por *tableta*)... S_for: *pretender* (co significado de 'finxir')...
gal	Palabras, trazos ou usos transferidos desde unha variedade do galego non estándar. No ámbito morfolóxico, flexión propia do galego popular. No ámbito léxico, palabra tirada dunha variedade do galego (popular ou culto). No ámbito gramatical, uso propio dalgunha variedade dialectal. No ámbito semántico, uso tirado dunha variedade do galego (galego popular ou culto)	M_gal L_gal G_gal S_gal	M_gal: *consegueu* (por *conseguiu*)... L_gal: *atrais* (por *atrás*)... G_gal: uso de *lle* por *lles*, cheísmo e teísmo S_gal: *mirar* (co valor de *ver*)...
hc	Hipercorreccións xeradas dentro do galego	L_hc G_hc	L_hc: hiperenxebrismos como *vaciña* (por *vacina*), *calqueira*... G_hc: colocación enclítica do pronome en vez de proclítica e uso de *che* por *te*
mix	Mesturas entre dúas palabras ou expresións	L_mix	L_mix: *indixesta* (*indixestión* + *inxesta*), *a saber Deus* (*a saber* + *sabe Deus*), *adedicar* (*adicar* + *dedicar*)...

Táboa 4. Continuación

or	Unicamente se emprega no ámbito gramatical para marcar os casos de pronomes enclíticos posibles na lingua oral, pero non recomendados na escrita formal	G_or	G_or: colocación enclítica do pronome posible na lingua oral, pero non recomendada na escrita formal
sp	Palabra, expresión, trazo ou uso transferidos desde o español. No ámbito ortográfico, grafía coincidente co español ou acentuación gráfica explicable por influencia desta lingua. No ámbito morfolóxico, flexión coincidente co español. No ámbito gramatical, colocación proclítica do pronome, así como ausencia ou adición de determinante, adición de preposición ou de pronome e substitución de determinante ou de pronome explicables por influencia do español. No ámbito léxico, unidades léxicas, acentuación de intensidade ou xénero transferidos desde o español. No ámbito semántico, calcos semánticos do español	O_sp M_sp G_sp L_sp S_sp	O_sp: *ahí* (por *aí*), *él*… M_sp: *anglosaxones*… G_sp: *levarse o diñeiro* (por *levar o diñeiro*)… L_sp: *actitud*, *élite*, *(a) leite* S_sp: *pobo* (por *vila* ou *aldea*)…
sp_adapt	No ámbito morfolóxico, flexión adaptada desde o español. No ámbito léxico, palabras ou expresións transferidas desde o español e adaptadas ao galego, ou mesturas entre elementos do galego e do español	M_spadapt L_spadapt	M_spadapt: *atraxo*, *supoñe*… L_spadapt: *semexante, ventaxes, chear*…

Nótese que, máis alá daquelas formas diverxentes do estándar cuxa orixe se nos escapa, existe un un uso restrinxido desta anotación. Isto é, non todas as formas para as que é posible ofrecer unha hipótese plausible relativa á súa orixe reciben un código deste tipo, senón só aquelas que se corresponden co exposto na Táboa 4: por exemplo, a marca de oralidade podería utilizarse, pero non se emprega, para etiquetar o uso da contracción *ca* en vez de *coa*. Son razóns de carácter práctico, derivadas da enorme complexidade do proceso de anotación, así como da procura da maior coherencia posible, as que explican este proceder. Por outro lado, debe terse en conta que o emprego das etiquetas é en certos casos bastante cauto. Así, por exemplo, en relación coa influencia do español na acentuación gráfica, só identificamos con O_sp os casos de adición e non os de omisión: a modo de ilustración, márcanse con este código os adverbios en *-mente* acentuados graficamente, pero non

as frecuentes omisións de acento na secuencia [u'i] (por exemplo *construir, incluir*, *ruido*), que tamén se poderían explicar polas diferenzas entre o sistema de acentuación do galego e do español (*vid*. o capítulo de López-Sández e Lorenzo-Herrera neste volume).

Lémbrese, por último, que, tal e como indicamos, un mesmo *token* pode recibir varios códigos explicativos. Así, por exemplo, os empregos cheístas van etiquetados tanto con G_hc como con G_gal, pois poden responder tanto a unha ultracorrección como a un uso dialectal. A introdución deste tipo de código realízase nun campo específico do mesmo formulario utilizado para a edición dos *tokens* presentado previamente (na Figura 18 pode encontrarse un exemplo).

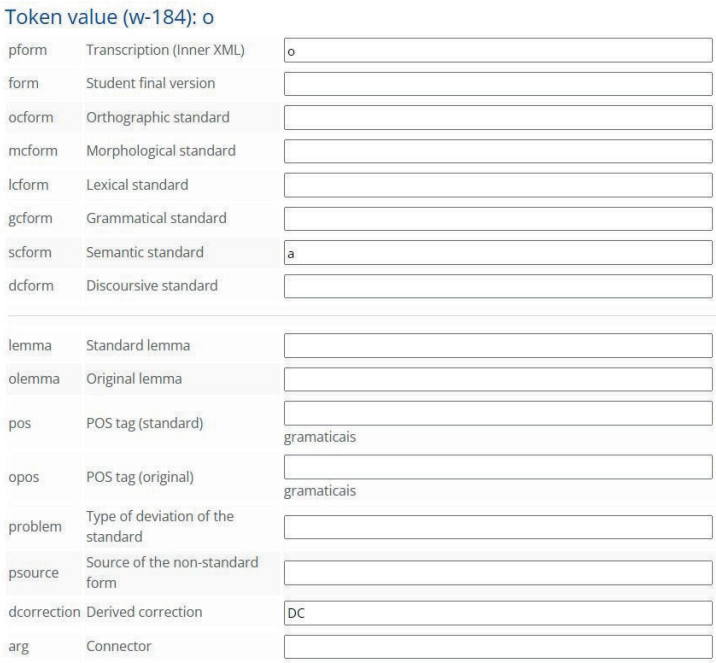
2.5.5. As correccións derivadas

En *CORTEGAL* podemos encontrar varios exemplos de formas normalizadas que non están asociadas a un código identificador de forma desviante, aínda que si ao código DC («derived correction»). Son dous os tipos de circunstancias en que utilizamos esta anotación.

En primeiro lugar, asignamos o código DC cando é necesario realizar un cambio no texto como consecuencia doutra estandarización levada a cabo neste, dun modo parcialmente similar ao que se propón para o corpus *ASK*. Neste corpus utilízase unha etiqueta específica para os chamados «agreement errors»: «errors following logically from, and triggered by, previous errors, the agreement itself being in accordance with the target language norm» (Tenfjord et al., 2006, p. 1822). Existen con todo algunhas diferenzas importantes entre o sistema empregado no corpus *ASK* e en *CORTEGAL*: no corpus *ASK* os «agreement errors» considéranse como unha subcategoría de erro ligada ás categorías «Puntuación», «Maiúsculas e minúsculas» e «Selección errónea de categoría morfosintáctica», mentres que en *CORTEGAL* a marca DC non se considera un código de «forma non estándar» e ademais esténdese a calquera caso en que sexa necesario realizar unha estandarización derivada doutra.

Así, por exemplo, ao estandarizar o calco semántico do español *pobo* por *vila* na frase «o pobo de Iago Aspas», é necesario substituír o artigo masculino polo feminino para evitar un problema de concordancia («a vila de Iago Aspas»). Agora ben, o/a estudante non comete ningún erro de concordancia, de tal modo que carecería de sentido asignarlle a *o* un código de desviación gramatical. Nese caso asignamos no determinante o código DC e a forma normalizada «a», sempre no mesmo nivel no que se realiza a estandarización que a orixina (neste caso no nivel semántico), tal e como se pode comprobar na Figura 21.

FIGURA 21. Anotación dunha corrección derivada



Deste, xeito, a estandarización de *pobo* por *vila* vai acompañada da modificación de *o* por *a* e ambos os cambios son visualizados simultaneamente na capa de estandarización semántica, tal e como se ve na Figura 22, onde está activada esta última. O determinante modificado aparece en azul, como *vila*, pero non leva asignado ningún código identificador dun tipo de desviación, senón simplemente a marca DC.

FIGURA 22. Visualización das formas con corrección derivada

Un exemplo do mesmo tipo atopámolo naqueles casos en que propomos a inclusión dun punto, o que conduce á necesidade de empregar maiúscula inicial na palabra seguinte. Así, por exemplo, no seguinte fragmento,

(9) Anteriormente, o turismo non estaba tan potenciado, habitualmente podemos observar numerosos anuncios por medios nos que se ofertan billetes de avión, barco etc. a bós prezos para que a xente se anime a viaxar e coñecer novos lugares

propoñemos substituír a coma que figura antes de *habitualmente* por un punto, o que obriga a utilizar maiúscula na primeira letra desta palabra, que pasa a etiquetarse con DC. Tamén se emprega o código DC, por exemplo, para a normalización das formas concordadas con palabras que reciben a etiqueta L_gen_su, a cal serve para indicar que se produce unha asignación de xénero non estándar a un substantivo. Sendo así, en (10), onde destacamos en cursiva as voces concordadas con *leite*,

(10) A leite *francesa* resulta mais *barata* (o cal quere dicir que costa uns poucos céntimos menos) *ca galega*

non parece moi apropiado asignar códigos de desviación gramatical a tales voces, que concordan adecuadamente co xénero que a persoa autora do texto lle atribúe a *leite*. Daquela, todas esas formas son normalizadas no nivel

léxico (o mesmo en que se etiqueta *leite*) e anotadas con DC, pero non reciben ningún outro código que identifique unha desviación e que se compute como tal (sobre o código L_gen_su, e a súa diferenciación con G_gen_su, que marca un problema de concordancia, *vid.* §2.5.6 e o capítulo de Álvarez de la Granja neste volume).

Outros moitos casos implican a anotación con DC: a estandarización dos pronomes proclíticos que deberían colocarse como enclíticos (o código de desviación gramatical asígnaselle ao pronome mal colocado, que se suprime, mentres que ao verbo se lle engade o pronome enclítico, acompañando esta modificación de DC); o engadido dunha preposición ao levar a cabo unha estandarización no nivel léxico ou semántico dun verbo que non rexe preposición por outro que si o fai; a inclusión dunha forma léxica no canto dun clítico proclítico cun referente irrecuperable no texto (o clítico anótase con código de desviación e suprímese, mentres que a forma léxica se etiqueta con DC) etc.

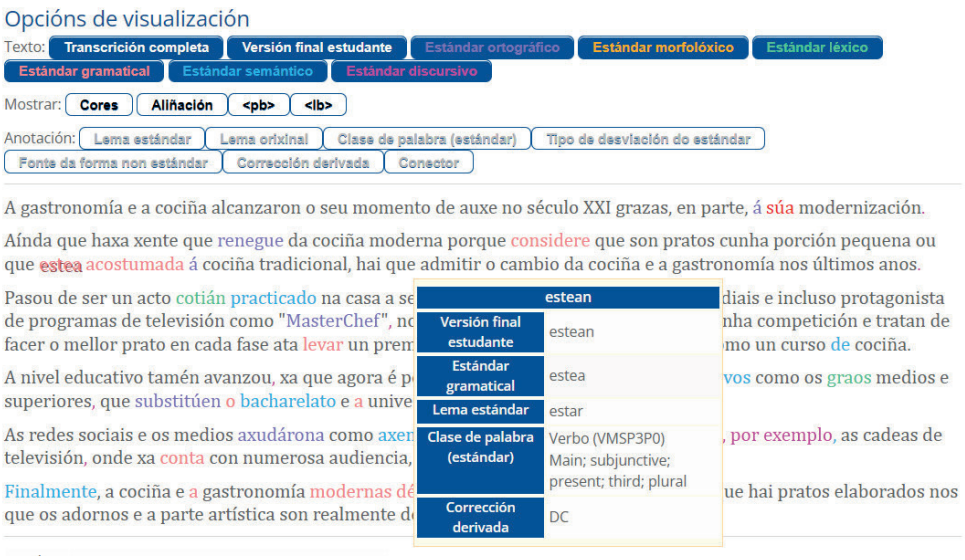
O segundo tipo de circunstancias en que se emprega DC é aquel en que un mesmo e único problema ten varias manifestacións no texto. En tales casos, asignamos código de forma non estándar só nunha desas manifestacións e DC nas restantes. Un exemplo deste tipo atopámolo nos problemas de concordancia, que moitas veces se materializan en varias palabras. Así sucede, por exemplo, nos seguintes fragmentos:

(11) Aínda que haxa xente que renege da cociña moderna porque *consideren* que son pratos cunha porción pequena ou que *estean acostumados* a cociña tradicional, hai que admitir o cambio da cociña e a gastronomía nos últimos anos
(12) O capitalismo e o afán de poder *desaparecería* polo feito de non necesit*alo*

Nas tres palabras destacadas en cursiva en (11) hai concordancia *ad sensum*, e as tres voces deberían ser estandarizadas, respectivamente, por *considere*, *estea* e *acostumada* para lograr a concordancia gramatical con *xente*. Un problema do mesmo tipo figura en (12), onde o verbo e o pronome destacados deberían estar en plural pola súa concordancia cunha construción coordinada. Á hora da anotación temos dúas opcións: colocar unha etiqueta G_num_su en cada palabra, de tal modo que no caso de (11) se computarían tres problemas diferentes de concordancia de número e no caso de (12) dous, ou ben colocar a etiqueta na primeira delas e simplemente modificar as demais, pero sen asignarlles código de forma non estándar (en *acostumada*, con todo, si se incluiría un código de desviación relativo á concordancia de xénero).

Aínda que nestes exemplos sería xustificable a anotación de todas as palabras, optamos por etiquetar como desviante só unha delas. A razón estriba fundamentalmente no cómputo de formas non estándar: ao noso entender, en casos como (11) e (12) é máis adecuado considerar que existe un único problema de concordancia (un problema na concordancia con *xente* e con *o capitalismo e o afán de poder*, respectivamente), que ten diferentes manifestacións no texto, antes que varios problemas computables de maneira independente. Dado que, en calquera caso, todas as formas con problemas de concordancia son estandarizadas, ao consultar o texto, todas elas, tanto as anotadas con código de desviación como as que non o están, destacarán convenientemente coa cor correspondente (neste caso salmón, ao tratarse dunha estandarización realizada no nivel gramatical). Ademais, os *tokens* modificados pero non etiquetados con código de forma non estándar irán acompañados da etiqueta DC, que dá conta de que responden á segunda, terceira... manifestación dun problema xa anotado previamente, tal e como se pode comprobar na Figura 23.

FIGURA 23. Visualización dunha corrección derivada nun caso de concordancia



Por razóns do mesmo tipo, utilizamos o procedemento indicado no caso dos incisos, conectores... que deberían estar demarcados por dúas comas no texto, pero en que ambas faltan, como sucede coa oración de xerundio que figura en (13):

(13) Grupos de homes, mulleres e nenos son obrigados construír empregando materiais tóxicos eses obxectos

Unicamente marcamos con código identificador de omisión de signo de puntuación (D_pm_om) a primeira das comas ausentes e incorporadas, mentres que a segunda coma se introduce exclusivamente coa marca DC. Consideramos que estamos ante un único problema (a ausencia de delimitación dun inciso) ao que lle corresponde unha dobre manifestación no texto e entendemos, en consecuencia, que debería computar unha soa vez. O mesmo sistema é empregado cando nun texto se deben incluír parénteses ou comiñas de apertura e peche.

2.5.6. Sobre a subxectividade da estandarización e da codificación en CORTEGAL

Como se sinala en moitos traballos (*vid.* por exemplo, Granger, 2003, p. 473; Díaz-Negrillo e Fernández Domínguez, 2006, p. 89; Reznicek et al., 2013, p. 104 ou Granger, 2017) a estandarización e a codificación das formas non estándar é un proceso complexo que encerra unha carga alta de subxectividade: «The subjectivity involved in identifying, correcting and classifying errors, which makes the process a time-consuming, effort-driven, and challenging task (...), is a characteristic of the error-tagging process» (Díaz-Bedmar, 2021, p. 95). A subxectividade na anotación de *CORTEGAL* faise presente de diferentes maneiras e con implicacións de maior ou menor relevancia.

Por un lado, en ocasións existen varias posibilidades para escoller *target hypothesis*, como no caso de *sin embargo* ou de *sen embargo*, que poden ser normalizados mediante diferentes sinónimos como *porén*, *con todo*, *no entanto*... En casos coma estes, optamos en xeral por unha mesma forma estandarizadora para todos os exemplos, seleccionada de acordo con criterios de frecuencia e adecuación, pero isto non implica necesariamente que unha mesma forma desviante se normalice sempre da mesma maneira, posto que o contexto ás veces é máis axeitado para unha *target hypothesis* que para outra (por exemplo, *conlevar* estandarízase ás veces con *implicar* e outras con *supoñer*).

Noutros casos, a opcionalidade na escolla de *target hypothesis* deriva da existencia de diferentes interpretacións ao respecto do que a persoa autora dos textos quería realmente dicir. Así, por exemplo, nunha das redaccións, falando de fútbol, dise que é un «deporte con moitos afiliados», cun emprego da palabra *afiliados* semanticamente inadecuada. Estandarizamos a voz por

seguidores, pero podería ser tamén normalizada mediante *socios*, que non porta o mesmo significado ca a palabra que finalmente seleccionamos.

Ás veces o contexto pode axudar a interpretar o sentido buscado pola persoa que escribe os textos. Así, nun exemplo como o seguinte,

(14) O consumismo e a productividade é un dos temas máis tratados na actualidade, debido a relación de dependencia que se establece entre elas, e *como* algún cambio nunha pode repercutir na outra

podemos deducir, pola frase introducida por *debido a*, que o valor de *como* (destacado por nós en cursiva) é causal, de modo que o normalizamos con *dado que*.

As persoas usuarias do corpus deben ter sempre en mente que as propostas de estandarización realizadas son en moitas ocasións unha simple escolla de entre varias opcións posibles. En calquera caso, tal e como indicamos, procuramos a confluencia na selección mediante o establecemento previo dunha forma estandarizadora naquelas formas desviantes do corpus recorrentes.

Nos dous tipos de casos sinalados, a subxectividade afecta exclusivamente á selección da forma normalizada, pero non á anotación do tipo de desviación do estándar. Noutros casos, a escolla atangue a esta. Nun exemplo como o anterior, (14), poderiamos en principio pensar que a presenza de *a* en vez de *á*-en «debido a relación de dependendencia...») ten tres interpretacións posibles: pode verse como un problema de acentuación, como ausencia da preposición ou como ausencia de artigo.

Agora ben, nun caso como o que acabamos de presentar, non podemos obviar as características xerais das persoas que crearon os textos: estamos tratando (cando menos na inmensa maioría dos casos) con estudantes que estudan galego e en galego desde primaria e resulta pouco esperable que haxa problemas á hora de usar esa estrutura, coincidente ademais co castelán (cambiando os complementos, parece pouco probable atopar *debido iso*, sen preposición, ou *debido a motivo indicado*, sen artigo), pero si é frecuente atopar problemas na acentuación (*vid*. o capítulo de López-Sández e Lorenzo-Herrera neste volume). Particularmente, a ausencia de acento na contracción do artigo definido feminino coa preposición *a* é moi habitual (300 casos en *CORTEGAL*, computando só a contracción co artigo feminino singular).

Pero outras veces, a escolla da anotación que se debe realizar ofrece bastantes máis dificultades: o adxectivo *sana*, que se rexistra 14 veces en *CORTEGAL* e cuxa correspondente forma estándar é *sa*, pode interpretarse como unha formación non estándar do feminino de *san* (seguindo o modelo de

folgazán, *folgazana*) e, daquela, como unha desviación morfolóxica (M_gen_su), ou como unha desviación léxica (L_w_su), pola escolla do castelanismo *sano, sana* en vez da unidade léxica estándar do galego *san, sa*.

En casos como os indicados, sobre todo nos máis dubidosos, como o de *sana*, preséntasenos unha dobre opción metodolóxica: escoller unha soa das anotacións ou anotar a forma diverxente con todos os códigos posibles. Malia o nesgo que isto pode supoñer, inclinámonos pola primeira das opcións: escoller un único código identificador do tipo de desviación. A razón fundamental é o feito de que empregar máis dunha etiqueta suporía non poder realizar de maneira doada o cómputo de desviacións do estándar presentes no corpus, identificables e computables precisamente a través dos códigos asignados, posto que unha mesma desviación estaría anotada con varias etiquetas.

Á hora de escoller a anotación que se vai realizar entre as varias opcións posibles, empregamos como criterios fundamentais a naturalidade e frecuencia do tipo de desviación. Así, tal e como sinalamos, resultaría moi estraña a omisión de preposición ou artigo en (14), de tal xeito que etiquetamos con O_ac_om (omisión de acento gráfico). No segundo exemplo, dado que no corpus hai máis casos de *san* que de *sano*, optamos por consideralo, en xeral, un problema morfolóxico. Con todo, tamén temos en conta, como criterio secundario, o contexto: se unha persoa que utiliza *sana* emprega no mesmo texto *sano*, interpretamos *sana* como unha desviación léxica. Do que acabamos de expoñer, dedúcese que unha mesma forma desviante pode ter ocasionalmente dúas anotacións diferentes en textos distintos, condicionadas pola información que se extrae do contexto.

A subxectividade da anotación vai implícita tamén na aplicación xeral dalgunhas etiquetas concretas. Por exemplo, a persoa anotadora debe decidir cando a selección inadecuada de xénero responde a un problema gramatical de concordancia (G_gen_su) e cando obedece a unha atribución de xénero non estándar a unha unidade léxica, en cuxo caso, a etiqueta sería léxica (L_gen_su). Así, por exemplo, consideramos que en «a leite francesa» hai un problema léxico, anotado sobre *leite*, posto que entendemos que se lle atribúe a esta palabra xénero feminino. Pola contra, en «a cultura xeralmente abre máis portas e non termina sendo tóxico» a anotación realízase sobre *tóxico* e non sobre *cultura*, dado que se considera que hai un problema de concordancia do adxectivo e non de asignación a *cultura* de xénero masculino. A razón do distinto tratamento obedece a criterios de frecuencia (hai en *CORTEGAL* 15 exemplos de uso en feminino de *leite*) e á existencia

dalgunha razón que xustifica a atribución dun xénero non estándar (o cognado do castelán *leche* é masculino). Decisións moi similares afectan á distinción entre erros ortográficos de acentuación (por exemplo, *ademaís*, considerado como un erro na colocación do acento gráfico: O_ac_wp) e erros léxicos de selección non estándar da sílaba tónica (por exemplo, *élite*, fronte ao galego estándar *elite*, considerado como un erro léxico: L_ac_su). As razóns son do mesmo tipo ca as indicadas no caso anterior (hai 7 exemplos de *élite* e existe o cognado español *élite*). Pode atoparse máis información sobre estas cuestións e sobre outras decisións do mesmo tipo no capítulo de Álvarez de la Granja neste volume.

Finalmente, debemos mencionar aqueles casos en que a opcionalidade da estandarización radica na posibilidade de normalizar elementos diferentes, o que pode implicar anotacións de distinto tipo.

Así, por exemplo, nun fragmento como este

(15) Pola contra noutros países que non teñen tantos recursos como por exemplo os países máis pobres de África, non producen nin consumen como os países europeos

poderiamos optar hipoteticamente por dúas estandarizacións (incluímos só as modificacións que afectan ao que nos interesa ilustrar):

(16) Pola contra, *outros* países que non teñen tantos recursos como por exemplo os países máis pobres de África, non producen nin consumen como os países europeos
(17) Pola contra noutros países que non teñen tantos recursos como, por exemplo, os países máis pobres de África, non *se produce* nin *se consume* como *nos* países europeos

No primeiro caso, o adxunto inicial «pasa a funcionar» na nova estrutura como suxeito. No segundo caso, modificamos a segunda parte do fragmento, mantendo o adxunto como tal e, dada a ausencia de suxeito, empregando unha estrutura impersoal.

En situacións deste tipo, o criterio xeral para a escolla da estandarización é o de optar pola forma normalizada máis sinxela e máis próxima ao texto da/do estudante, que neste caso corresponde a (16).

Así pois, e a maneira de conclusión do exposto, aínda que a subxectividade na estandarización e etiquetaxe é inevitable, procuramos utilizar algúns criterios que aseguren a maior homoxeneidade posible: por un lado, descartamos sempre aquelas interpretacións dunha desviación que son posibles pero estrañas ou forzadas. Buscamos sempre a interpretación máis natural

tendo en mente as características xerais das persoas que elaboraron os textos, as cales, teñan o galego como L1 ou L2, estudan e coñecen a lingua, cando menos na gran maioría dos casos, desde a educación primaria. Ademais, o criterio da frecuencia é fundamental, tanto para a escolla das formas normalizadas como para determinar a anotación que se realiza. Tamén temos en conta a proximidade ao texto orixinal e a información que este nos pode proporcionar, tanto á hora de estandarizar como de anotar. Debemos indicar, ademais, que a posibilidade que nos ofrece *TEITOK* de editar conxuntamente varias formas desviantes (por exemplo, podemos estandarizar e etiquetar conxuntamente todos os exemplos de *aunque* presentes en *CORTEGAL*), non só facilita o proceso, senón que contribúe a diminuír o número de erros e a lograr maior coherencia.

Os criterios expostos están recollidos no manual de estandarización e etiquetaxe, xa mencionado en §2.5.4, que estivo a disposición das persoas responsables desas dúas tarefas ao longo de todo o proceso. Antes da aplicación das indicacións do manual no corpus, realizáronse probas co obxectivo de adestrar as persoas encargadas das dúas tarefas, todas elas lingüistas bilingües en galego e castelán. O manual contén os criterios xerais xa mencionados, así como os códigos empregados na anotación, con exemplos, explicacións e observacións adicionais sobre o seu emprego, e tres apéndices para a estandarización de aspectos especialmente problemáticos: uso dos signos de puntuación, uso do xerundio e uso da preposición con obxecto directo.

A partir da análise dunha mostra de textos de *CORTEGAL* e das propostas realizadas para outros corpus con anotación informatizada de erros elaborouse unha versión piloto para a estandarización e etiquetaxe, que se testou con outra pequena selección de textos. Tendo en conta os problemas detectados, elaborouse unha segunda versión do manual, con algunhas modificacións na codificación provisoria e co engadido de novas explicacións e aclaracións. Esta nova versión avaliouse á vista das dificultades de aplicación e do acordo intra- e interanotadoras/es tanto para a estandarización como para a etiquetaxe dun conxunto de textos, resultando unha terceira versión, coa que se deu por rematada a pilotaxe. Con todo, esta versión sufriu tamén varias modificacións ao longo do proceso de normalización e codificación dos 1000 textos, dada a necesidade de realizar algúns axustes e aclaracións. Para asegurar a maior homoxeneidade posible, a estandarización e etiquetaxe de todos os textos foi revisada por unha das persoas responsables do corpus, a cal, unha vez rematado o proceso, tamén levou a cabo a revisión da aplicación coherente dos códigos a través da análise das liñas de concordancia proporcionadas polo buscador. Desta revisión derivaron novos cambios, para

lograr unha maior consistencia na etiquetaxe, que conduciron á versión final do manual, que é a que se encontra dispoñible, como xa foi indicado, na páxina do proxecto. Por suposto, despois de cada revisión do manual, levouse a cabo tamén a revisión da estandarización e da etiquetaxe realizadas ata ese momento para aplicar os novos criterios, asegurando deste xeito a maior coherencia posible ao longo de todo o corpus.

A pesar das continuas revisións, son inevitables algunhas inconsistencias, que corriximos a medida que detectamos. O corpus está aberto tamén á participación das persoas usuarias, que poden facer chegar as súas suxestións e indicacións ás/aos responsables de *CORTEGAL*. En calquera caso, tales inconsistencias non deben confundirse coas interpretacións, inevitables na estandarización e etiquetaxe dos corpus de aprendentes: non podemos considerar que «the error annotation which is present in a corpus is the 'truth' or 'correct analysis' instead of just one among many interpretations» (Lüdeling e Hirschmann, 2015, p. 142).

2.6. Asignación de lemas e categorías gramaticais (estándar e orixinais)

A lematización realízase automaticamente en *TEITOK* a través da versión 4.2 de *FreeLing* en lingua galega. O proceso de lematización con *FreeLing* implica a asignación de lema estándar e categoría gramatical a todos os *tokens*, incluídos os signos de puntuación (no caso das contraccións e verbos con pronomes enclíticos, só aos *dtokens*, vid. §2.2). A anotación morfosintáctica emprega etiquetas *EAGLES* (Leech e Wilson, 1996).

A lematización realízase unha vez normalizadas e etiquetadas as formas non estándar e sempre sobre a versión lexicamente estandarizada, que implica tamén a estandarización ortográfica e morfolóxica, tal e como vimos en §2.5.3. Escóllese a dimensión léxica, pois é o nivel máis próximo ao texto orixinal que asegura a identificación e lematización adecuada dos *tokens* por parte de *FreeLing*.

Deste xeito, a lematización dun texto como (18), en que se destacan en cursiva as formas que deben ser estandarizadas na dimensión ortográfica (*conflictos*), na morfolóxica (*corporales*) e na léxica (*sen embargo, tareas*),

(18) Na época de adolescencia é unha época chea de *conflictos* xa que é unha etapa dura nosas vidas, e cando somos tan novos e inexpertos necesitamos o apoio de figuras paternas e maternas, *sen embargo* non é unha *tarea* doada xa que é a idade con máis cambios de humor e incomprensión por parte dos nenos xa que está chea de cambios, tanto *corporales* como na mentalidade

realízase sobre (19), unha vez realizadas as correspondentes modificacións, posto que formas como *corporales*, *tarea* ou *conflictos* non son recoñecidas por *FreeLing*.

(19) Na época de adolescencia é unha época chea de *conflitos* xa que é unha etapa dura nosas vidas, e cando somos tan novos e inexpertos necesitamos o apoio de figuras paternas e maternas, *porén* non é unha *tarefa* doada xa que é a idade con máis cambios de humor e incomprensión por parte dos nenos xa que está chea de cambios, tanto *corporais* como na mentalidade

Pola contra, o lematizador non alcanza as estandarizacións que se realizan no nivel gramatical, semántico ou discursivo. Así, por exemplo, no texto anterior substitúese no nivel discursivo a coma que precede a *sen embargo* / *porén* por un punto, pero o lema e categoría gramaticais asignados son «,» e «Fc» e non «.» e «Fp». De igual xeito, a contracción *das*, incluída no nivel gramatical despois de «unha etapa dura», non é obxecto de lematización, como tampouco o é ningún outro elemento incorporado no texto polas persoas encargadas da estandarización cando hai algunha omisión.

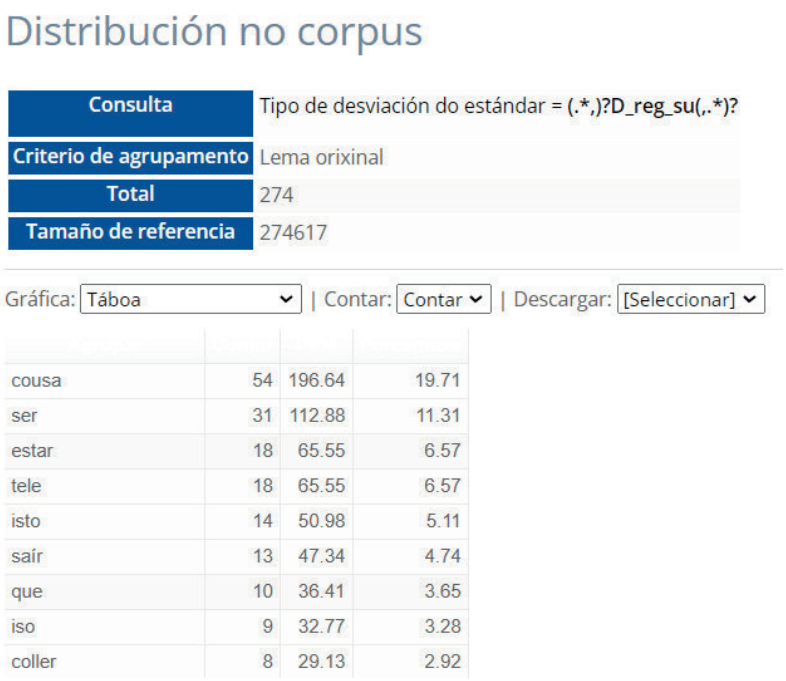
Un aspecto metodolóxico importante de *CORTEGAL* é a asignación manual, na maior parte das formas estandarizadas lexicamente, dun segundo lema, chamado lema orixinal e, nunhas poucas destas formas, dunha segunda categoría gramatical, denominada clase gramatical orixinal. Así, a forma *tarea* que figura en (18) recibe dous lemas: o lema estándar *tarefa*, asignado automaticamente mediante *FreeLing*, e o lema orixinal *tarea*, asignado manualmente. O lema orixinal é a forma de cita que representa a voz escrita pola/polo estudante (unha vez estandarizada, se é o caso, ortográfica e morfoloxicamente), así como calquera outra variante flexiva que poida haber no corpus. Así, outros exemplos de lemas orixinais introducidos en *CORTEGAL* son *sen embargo*, *abuelo* (común a *abuela*, *abuelo* e *abuelos,* as tres formas recollidas no corpus), *conlevar* (representante de *conleva*, *conlevan*, *conlevar*, *conlevaron* e *conlevou*), *novidoso* (asignado a *novidoso*, *novidosa*, *novidosos*, *novidosas* e ao lapsus ortográfico *novidodos*), *cachivache* (atribuído a *cachibaches*) etc. No caso de que non se atribúa ningún lema orixinal manualmente, este coincide por defecto co lema estándar. Tal e como se indicou, o lema orixinal só se lles asigna manualmente ás formas que son estandarizadas no nivel léxico, concretamente a aquelas que portan un código L_w_su (substitución dunha unidade léxica estándar por unha non estándar, con diferenzas que van máis alá da acentuación de intensidade ou do xénero, como *abuela* por *avoa*) ou un código L_ac_su (substitución dunha unidade léxica estándar por unha non estándar, con diferenzas de acentua-

ción de intensidade, como *élite*). Deste xeito, nas formas etiquetadas con códigos diferentes a L_w_su e L_ac_su, así como nas voces que non reciben ningún tipo de modificación, o lema orixinal coincide co estándar.

A categoría gramatical orixinal introdúcese tamén manualmente cando existe algunha diferenza entre a clase ou subclase de palabras da forma escrita pola/polo estudante e da forma estandarizada no nivel léxico. O exemplo máis claro é o das voces etiquetadas con L_gen_su, en que se produce a asignación dun xénero non estándar a determinadas palabras, tal e como vimos en §2.5.5 e §2.5.6. Deste xeito, a palabra *leite* na secuencia «a leite» ten atribuído como clase de palabra estándar NCMS000 (substantivo común masculino singular), pero como clase de palabra orixinal NCFS000 (substantivo común feminino singular). Máis alá das formas etiquetadas con L_gen_su, podemos atopar outros casos de asignación manual de clase de palabra orixinal: nótese, por exemplo, a diferenza categorial entre o lema orixinal *aunque*, conxunción, e o lema estándar *aínda que*, adverbio + conxunción (*vid.* o indicado ao final deste apartado sobre o tratamento das expresións complexas). Como sucedía cos lemas, a asignación de clase de palabra orixinal só se realiza nas palabras estandarizadas no nivel léxico, de maneira que as estandarizacións que se realicen no nivel gramatical, semántico ou discursivo e que impliquen cambio categorial non implican a atribución manual dunha clase de palabra orixinal. Se non se atribúe ningunha categoría gramatical manualmente, esta coincide coa estándar, de xeito que naqueles casos en que a estandarización non implica mudanza categorial, así como nos *tokens* que non se modifican, a clase de palabra orixinal coincide coa estándar, atribuída por *FreeLing*.

A asignación de lema orixinal é especialmente útil na análise dos datos do corpus, sobre todo cando interesa realizar consultas sobre frecuencia tendo os lemas como criterio de agrupamento. Mediante as consultas sobre frecuencia que ofrece o buscador de *TEITOK*, é posible obter a distribución dalgún elemento (por exemplo, un código de anotación das formas non estándar) organizado de acordo con algún criterio (por exemplo, lemas). Así, por exemplo, se interesa obter unha listaxe lematizada das formas codificadas con D_reg_su, pode realizarse unha busca das formas etiquetadas con este código e posteriormente seleccionar como criterio organizador os lemas estándar ou ben os lemas orixinais. Deste xeito, obtense unha táboa coa listaxe de lemas, tal e como se ve na Figura 24 (ofrecemos un fragmento).

FIGURA 24. Frecuencia de D_reg_su por lema orixinal (fragmento da táboa xerada por *TEITOK*)



Distribución no corpus

Consulta	Tipo de desviación do estándar = (.*,)?D_reg_su(,.*)?
Criterio de agrupamento	Lema orixinal
Total	274
Tamaño de referencia	274617

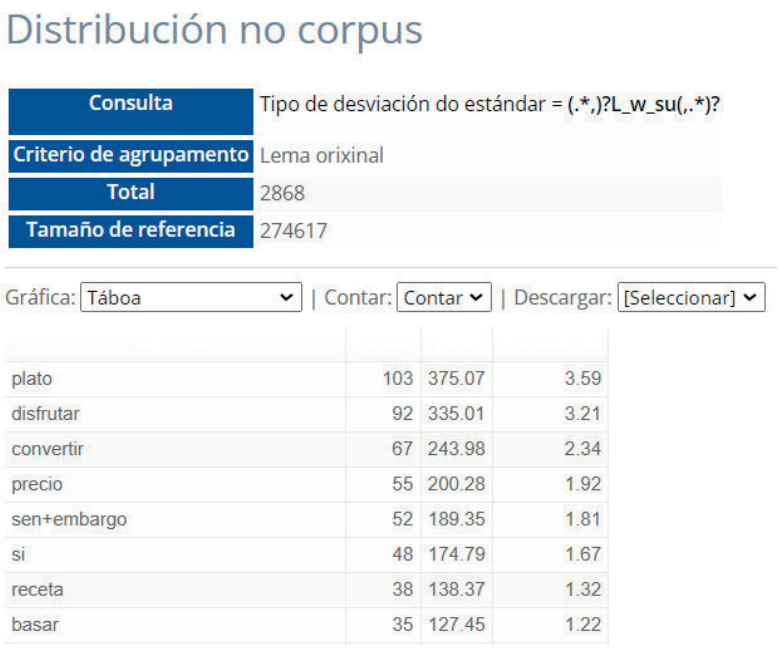
Gráfica: Táboa | Contar: Contar | Descargar: [Seleccionar]

cousa	54	196.64	19.71
ser	31	112.88	11.31
estar	18	65.55	6.57
tele	18	65.55	6.57
isto	14	50.98	5.11
saír	13	47.34	4.74
que	10	36.41	3.65
iso	9	32.77	3.28
coller	8	29.13	2.92

Agora ben, se se seleccionan os lemas orixinais obteremos unha listaxe máis próxima ao texto da/do estudante (por exemplo, teremos nesa listaxe *asqueroso*, correspondente á voz *asquerosa* escrita pola/o estudante). Se non houbese lema orixinal e só contásemos coa posibilidade de escoller como criterio organizador o lema estándar, na listaxe aparecería *noxento*, pois o castelanismo *asquerosa* foi estandarizado no nivel léxico mediante *noxenta*, de xeito que foi esta a voz que se lematizou.

O interese do lema orixinal é obvio cando o que se pretende estudar son as formas lexicamente non estándar (por exemplo, as etiquetadas con L_w_su), porque só deste xeito podemos obter unha listaxe lematizada das formas non estándar escritas polo alumnado, tal e como vemos na Figura 25 (fragmento). Por suposto, se só tivésemos lemas estándar, isto non sería posible, porque a táboa só incluiría estes últimos (Figura 26, fragmento).

FIGURA 25. Frecuencia de L_w_su por lema orixinal (fragmento da táboa xerada por *TEITOK*)



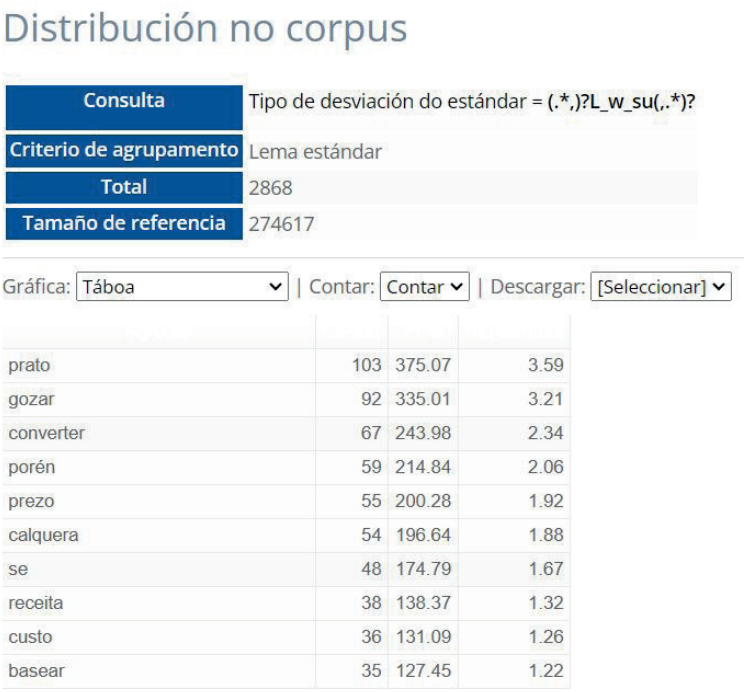
Distribución no corpus

Consulta	Tipo de desviación do estándar = (.*,)?L_w_su(,.*)?
Criterio de agrupamento	Lema orixinal
Total	2868
Tamaño de referencia	274617

Gráfica: Táboa | Contar: Contar | Descargar: [Seleccionar]

plato	103	375.07	3.59
disfrutar	92	335.01	3.21
convertir	67	243.98	2.34
precio	55	200.28	1.92
sen+embargo	52	189.35	1.81
si	48	174.79	1.67
receta	38	138.37	1.32
basar	35	127.45	1.22

FIGURA 26. Frecuencia de L_w_su por lema estándar (fragmento da táboa xerada por *TEITOK*)



Distribución no corpus

Consulta	Tipo de desviación do estándar = (.*,)?L_w_su(,.*)?
Criterio de agrupamento	Lema estándar
Total	2868
Tamaño de referencia	274617

Gráfica: Táboa | Contar: Contar | Descargar: [Seleccionar]

prato	103	375.07	3.59
gozar	92	335.01	3.21
converter	67	243.98	2.34
porén	59	214.84	2.06
prezo	55	200.28	1.92
calquera	54	196.64	1.88
se	48	174.79	1.67
receita	38	138.37	1.32
custo	36	131.09	1.26
basear	35	127.45	1.22

Por outro lado, e dado que o lematizador empregado non identifica como unidades as expresións complexas lexicalizadas, cada un dos seus constituíntes é lematizado de maneira independente. Así, por exemplo, o conector *pola contra* non constitúe unha unidade recoñecida como tal por *FreeLing*, de maneira que cada constituínte ten un lema e unha categoría gramatical asignados de maneira independente (preposición *por*, artigo determinado *o*, substantivo *contra*). Unicamente no caso de que sexa necesaria unha estandarización conxunta dunha expresión, levamos a cabo a súa unificación manual mediante a función *merge* que ofrece *TEITOK*, tal e como indicamos en §2.5.4, pero neste caso mantemos tamén a asignación de lemas e categorías gramaticais para cada compoñente, separados desta volta mediante o signo «+»: así, por exemplo, a locución *encher o buche*, que debe unificarse, posto que é substituída por *saciarse* no nivel discursivo ao non pertencer a un rexistro formal, ten como lema orixinal *encher* + *o* + *buche*, como lema estándar *encher* + *o* + *boche* e como clase de palabras (estándar e orixinal, pois non varía) VMN0000+DA0MS0+NCMS000 (infinitivo + artigo determinado masculino singular + nome común masculino singular).

2.7. Clasificación dos conectores que vinculan enunciados

TEITOK é unha plataforma flexible, que permite anotacións de moi diferente tipo e a incorporación destas en calquera momento. Consideramos que, dado que o estudo dos aspectos discursivos en lingua galega non recibiu ata o de agora demasiada atención, sería interesante a anotación no corpus dalgúns elementos con relevancia no nivel textual (máis alá das desviacións) para promover análises neste ámbito. Con todo, a complexidade e a esixencia do resto das anotacións determinou que esta tarefa se limitase ata o de agora a unha anotación sinxela daqueles conectores que teñen como función a vinculación entre enunciados (excluíndo vínculos temporais relativos á ordenación cronolóxica dos feitos na realidade[9]), aínda que a nosa intención é engadir no futuro novas anotacións discursivas.

[9] Así, por exemplo, non se anotaría o conector *despois*, nun uso como o que atopamos en «Os anos de posguerra e de franquismo foron anos de pobreza; a xente comía o que se podía permitir cociñar na súa casa. *Despois*, coa recuperación económica, a xente comezou a frecuentar máis os restaurantes». Si se anotaría *despois*, pola contra, nestoutro fragmento, posto que neste caso o conector serve para organizar as ideas do discurso: «Polo tanto hai distintos sectores de produción. Por exemplo, no sector de produción téxtil, vese condicionado pola moda de cada época. A nós tocounos vivir nunha época no que a moda ten un papel moi activo na nosa sociedade. Entón unha cousa dependerá da outra, dependendo das tendencias de cada momento a producción será distinta e as persoas beneficiadas tamén. *Despois* hai sectores como os da alimentación, que non dependen do mismo xeito da poboación, porque todo o mundo ten que comer, aínda que sexa o máis básico».

De acordo co indicado, anótanse en *CORTEGAL* os conectores que serven para unir enunciados, pero non os que establecen vínculos intraoracionais. Deste xeito, son etiquetados os elementos destacados en (20) e (21), pero non en (22) e (23):

(20) O consumo consiste na acción que realizamos todos de adquirir algo a cambio dun determinado prezo (capital). *Por outra banda*, ese algo que adquirimos é o produto
(21) Sen consumo, non hai produción, porque a demanda é inexistente, e se non hai produción, pois non pode haber consumo.
Pero o que parece fácil, se o analizamos profundamente, non o é
(22) Como se relata no texto, o consumo é unha necesidade e eso é moi certo, posto que é necesario para mercar a comida de cada día, *pero por outra banda* tamén di que serve para obter a máxima satisfacción
(23) Moitos dirán que tan só é unha moda, *pero* é necesario sinalar que a cociña e a gastronomía cambiáronlle a vida a moitas persoas

Os conectores textuais foron clasificados en 13 grupos de acordo co valor que achegan. Os códigos, valores e exemplos amósanse na Táboa 5:

Táboa 5. Códigos empregados para a anotación dos conectores que vinculan enunciados

Código	**Valor**	**Exemplos**
ad	Introducen un elemento que se suma ao dito anteriormente na argumentación, mesmo reforzándoo.	*ademais, e, é máis, tamén...*
caus	Introducen a causa do expresado no enunciado ou enunciados previos	*ao fin e ao cabo, pois, porque...*
coment	Introducen un comentario sobre o expresado anteriormente	*ben, pois, pois ben...*
concl	Introducen o resultado ao que se chega a partir do exposto previamente ou ben o seu resumo	*en conclusión, en definitiva, en resumo...*
confirm	Introducen un argumento, explicación... que serve para confirmar o que se acaba de dicir	*efectivamente, en efecto, si*
cons	Introducen a consecuencia do expresado no enunciado ou enunciados previos	*así, entón, por conseguinte...*
contraarg	Introducen un argumento ou explicación que se contrapón dalgún xeito ao que se acaba de indicar	*aínda así, agora ben, en cambio...*
disjunc	Introducen unha alternativa ao que se acaba de expoñer	*ou*
exempl	Introducen un exemplo ou concreción do que se acaba de indicar	*así, concretamente, por exemplo...*

Táboa 5. Continuación

fin	Introducen a finalidade do expresado no enunciado ou enunciados previos	*para*
ord_sp	Organizan o discurso distribuíndo os feitos expostos en diferentes apartados mediante elementos con valor espacial	*por outra banda, por unha parte, por un lado…*
ord_t	Organizan o discurso indicando a posición que ocupa cada elemento no proceso argumentativo ou expositivo do texto mediante elementos con valor temporal	*finalmente, por último, primeiramente…*
reform	Explican ou presentan doutra forma o que se acaba de expoñer	*é dicir, explícome, isto é*

Para levar a cabo a anotación destes conectores da maneira máis coherente posible, foi necesario tomar unha serie de decisións relativas tanto á delimitación da unidade «enunciado» como do concepto «conector». Estas decisións, que por razóns de espazo non poden ser comentadas aquí, están recollidas nun manual de anotación, que está dispoñible na páxina do proxecto (http://ilg.usc.gal/cortegal/downloads/conectores.pdf). Unicamente nos gustaría sinalar o feito de que a clasificación dos conectores se realiza atendendo ao valor que se lles atribúe por defecto[10], con independencia de que, no contexto en que figuran, o seu uso non sexa adecuado. Así, por exemplo, o conector *pero* que rexistramos no fragmento recollido en (24) é clasificado como contraargumentativo, por moito que no contexto non pareza haber ningún tipo de contraargumentación (polo que se anota con S_w_su e se substitúe por *e*).

(24) Na actualidade a gastronomía e a cociña acadaron moita popularidade. A gastronomía dos diferentes países é moi abundante pero tamén moi diferentes entre sí. *Pero* na actualidade son un factor clave á hora de visitar determinados países

Constitúen excepción os casos en que existen razóns para supoñer que estamos, non ante un mal uso, senón ante unha atribución errónea de significado, como sucede con *porén*, utilizado en varios textos con valor consecutivo, de tal xeito que lle atribuímos en tales textos o código *cons*. En (25) atopamos un exemplo deste tipo:

(25) Se a sociedade deixase de consumir produtos, as fábricas deixarían de producir e produciríase unha xigantesca crise económica, debido a que moitas persoas quedarían sen traballo, non recibirían un soldo, e non poderían consumir. *Porén*, Podemos dicir que é «unha pescadilla que se morde a cola»

[10] Por suposto, se o conector é polisémico, como sucede, por exemplo, con *así*, analizamos o contexto para atribuírlle o código que lle corresponde.

Finalmente, gustaríanos indicar que, no caso dos conectores pluriverbais (*así pois*, *por outra banda*, *sen embargo*...), o código asígnaselle a un dos elementos da unidade (aquel que parece achegar unha maior carga semántica; nos casos anteriores *pois*, *banda* e *embargo*), pero non levamos a cabo ningún tipo de fusión mediante *merge*, pois, como indicamos en §2.5.4, esta opción dá lugar a algúns problemas de recuperación e de etiquetaxe.

3. Principais resultados

3.1. Caracterización xeral dos textos. Análise cuantitativa

O número total de palabras dos textos é de 224.424, de acordo cos datos que ofrece *Dcontado*, que tal e como indicamos en §2.6, non computa os signos de puntuación e contabiliza as contraccións e as combinacións de verbo e pronome enclítico como unha soa unidade. Dado que o corpus está conformado por 1000 textos, a media de palabras por texto é de 224,42. Do total de palabras, 202.645 corresponden aos exames da convocatoria de xuño e 21.779 aos de setembro, de tal xeito que os 898 exames da primeira convocatoria teñen unha media de 225,66 palabras por texto e os 102 exames da segunda convocatoria unha media de 213,52, inferior á anterior en algo máis de 12 palabras. O texto cun menor número de palabras contén 51 (semella un texto que quedou sen rematar) e o máis longo ten 613. A desviación estándar é de 52,74.

Con respecto ao número de lemas ou palabras distintas, establecido tamén de acordo cos datos obtidos de *Dcontado*, este é de 113.348, de xeito que temos unha media de 113,35 lemas por texto. Os lemas distribúense entre 102.593 en xuño e 10.755 en setembro, o que supón unha media de 114,25 lemas en xuño e 105,44 en setembro, cunha diferenza de 8,81. O rango sitúase entre un mínimo de 37 e un máximo de 270 lemas, nos mesmos textos que servían para establecer o rango de palabras. A desviación estándar é de 22,69.

Para obter a densidade léxica, dividimos o número de lemas entre o número de palabras[11]. A densidade léxica media é de 0,51, cunha pequena diferenza entre os textos de xuño (0,51) e os de setembro (0,50), malia que, en principio, a razón entre *types* e *tokens* adoita diminuír a medida que aumenta a lonxitude do texto (Capsada Blanch e Torruella Casañas, 2017, p. 351) (e lembremos que os textos de xuño son máis longos ca os de setembro).

[11] Somos conscientes de que esta medida de densidade léxica, coñecida como TTR (*vid.* Capsada Blanch e Torruella Casañas, 2017), é moi simple e non ten en conta aspectos relevantes, como a extensión dos textos, pero razóns prácticas determinaron a escolla deste sistema de cálculo. En calquera caso, os datos e os textos están á disposición de calquera persoa que desexe realizar unha medición máis axustada.

O rango sitúase entre 0,35 e 0,73 (no texto máis breve, de 51 palabras e 37 lemas), sendo a desviación estándar de 0,05.

Lémbrese en calquera caso as precaucións indicadas en §2.6 sobre o cómputo de lemas e, daquela, da densidade léxica.

No que se refire ao número de enunciados, *DContado* contabiliza como tales as unidades que van de punto a punto. A media no total dos textos é de 9,03, sendo a diferenza entre xuño e setembro de máis dun enunciado (9,15 en xuño e 8,06 en setembro). O rango sitúase entre 2 enunciados en 3 textos e 29 nun (non se trata do texto máis longo, senón dun de 356 palabras). A desviación estándar é de 3,23.

Con respecto á media de palabras por enunciado, esta é de 26,96 e a diferenza entre xuño (26,66) e setembro (29,51) é de case tres palabras, pero desta volta a cifra máis alta corresponde á segunda convocatoria. O rango establécese entre 10,05 e 98,50 (nun dos textos con só dous enunciados), sendo a desviación estándar de 8,88.

A media de palabras do enunciado máis longo é de 47,51, e é tamén en setembro onde atopamos unha media máis alta, con 48,59 fronte a 47,39 en xuño, cunha diferenza de algo máis dunha palabra. O rango vai de 17 a 146 e a desviación estándar é de 15,91. E, con respecto ao enunciado máis curto, a media sitúase en 12,28, con 11,94 palabras en xuño e 15,32 en setembro. O rango vai de 1 (en 7 textos, estando o enunciado constituído en varios deles pola palabra *non*) a 93 nun texto. A desviación estándar é de 6,49.

En relación co número de parágrafos, a media é de 4,30, algo superior en xuño (4,32) ca en setembro (4,07). Un total de 44 textos están constituídos por un só parágrafo, mentres que o texto con maior número de parágrafos ten 12. A desviación estándar, pola súa parte, é de 1,60.

Finalmente, a media de enunciados por parágrafo é de 2,41, sendo algo máis alta na primeira convocatoria (2,42) ca na segunda (2,33). O rango sitúase entre un só enunciado por parágrafo en 40 textos e 11 en 5 textos, sendo a desviación estándar de 1,39.

Os datos presentados recóllense na Táboa 6:

Táboa 6. Datos cuantitativos sobre os textos de *CORTEGAL*

Media total	Media xuño	Media setembro	Desviación estándar	Rango
Palabras				
224,42	225,66	213,52	52,74	51-613
Lemas				
113,35	114,25	105,44	22,69	37-270
Densidade léxica				
0,51	0,51	0,50	0,05	0,35-0,73
Enunciados				
9,03	9,15	8,06	3,23	2-29
Palabras por enunciado				
26,96	26,66	29,51	8,88	10,05-98,50
Palabras do enunciado máis longo				
47,51	47,39	48,59	15,91	17-146
Palabras do enunciado máis curto				
12,28	11,94	15,32	6,49	1-93
Parágrafos				
4,30	4,32	4,07	1,60	1-12
Enunciados por parágrafo				
2,41	2,42	2,33	1,39	1-11

Tal e como se pode observar na táboa, os textos da convocatoria de xuño, que, de acordo co que vimos en §2.1, teñen unhas cualificacións considerablemente máis altas ca os da convocatoria de setembro (6,93 fronte a 5,21), son textos algo máis longos ca estes últimos. O tamaño medio dos textos (225,67 palabras) sitúase, curiosamente, na media dos dous extremos indicados na proba (solicítanse textos de entre 200 e 250 palabras). O alumnado de setembro fai menos enunciados pero algo máis longos, de 29,51 palabras, fronte ás 26,66 do alumnado de xuño, e estes enunciados distribúense nunha media total de 4,30 parágrafos e 2,41 enunciados por parágrafo, sendo ambas as medias bastante similares nas dúas convocatorias. Con todo, o rango da media de enunciados por parágrafo que máis se repite é o que se sitúa entre 1 e 1,99 (con 434 textos), de tal xeito que un 43% dos textos da primeira convocatoria teñen unha media de enunciados por parágrafo inferior a 2 e esa porcentaxe elévase ao 48% na convocatoria de setembro. De feito, o 7,84% dos textos de setembro teñen un único enunciado por parágrafo, mentres que en xuño esa porcentaxe baixa ao 3,56%.

3.2. *Tokens* e clases de palabras

O número total de *tokens* que contabiliza o corpus é de 274.617, pero debe terse en conta que *TEITOK* computa tamén as formas engadidas polas persoas editoras no proceso de estandarización. Se nos limitamos ás formas que escribiu a/o estudante, que son as únicas que se lematizan e as únicas ás que se lles asigna categoría gramatical, temos un total de 268.595 *tokens*, cuxa distribución por categorías se recolle na Táboa 7. Debe sinalarse que, cando creamos un *token* complexo mediante *merge*, *TEITOK* compútao na clase do seu primeiro compoñente, de tal xeito que optamos por crear a categoría «Expresión complexa», e contabilizar conxuntamente todas estas formas constituídas por varios elementos, restándoas do cómputo do compoñente inicial[12].

Táboa 7. Distribución dos *tokens* en clases de palabras

Categoría	**Número de *tokens***	**Porcentaxe sobre o total**
Substantivo	52406	19,51%
Determinante	44452	16,55%
Verbo	40936	15,24%
Preposición	35337	13,16%
Signo de puntuación	25004	9,31%
Conxunción	19420	7,23%
Pronome	18292	6,81%
Adxectivo	15945	5,94%
Adverbio	15896	5,92%
Numeral	595	0,22%
Expresión complexa	302	0,11%
Interxección	10	0,00%
Total	*268.595*	

3.3. As formas non estándar

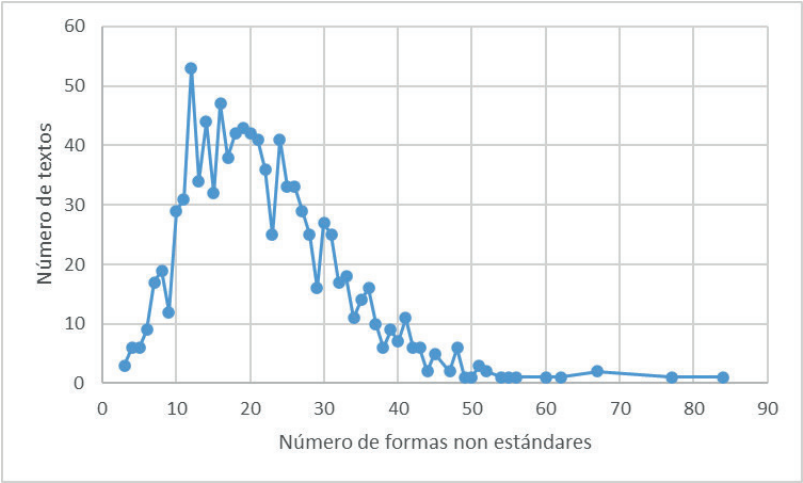
Tal e como indicamos en §2.5.4, asignámoslles a todas as formas non estándar un código que describe o tipo de desviación e, nalgúns casos, unha

[12] Por tal motivo, nalgunhas categorías, a cifra que ofrecemos non coincide exactamente coa que indica *TEITOK*.

segunda etiqueta que indica a súa orixe. Centrarémonos en primeiro lugar na codificación descritiva, á que corresponden un total de 23.169 etiquetas. Isto quere dicir que a media de códigos que sinalan algún tipo de desviación do estándar é de 23,17 por texto. A cifra total e a media indicadas non coinciden exactamente co número total de *tokens* anotados no corpus e coa media de *tokens* etiquetados por texto, respectivamente, posto que a unha mesma forma poden asignárselle varios códigos e porque en varias ocasións os códigos non etiquetan formas do corpus, senón que identifican omisións de palabras, que nós introducimos en *CORTEGAL*. Se consideramos os *tokens* anotados así como os creados por nós, asociados a códigos de desviación do estándar, o número total de formas anotadas é de 21.940, cifra á que hai que sumar as 513 secuencias ou palabras anotadas en *stand-off*, o que dá como resultado un total de 22.453 *tokens* ou secuencias con códigos de forma non estándar e unha media de 22,45 por texto. Se comparamos os datos das convocatorias de xuño e de setembro, observamos que a media é máis alta nesta segunda oportunidade, o que resulta coherente co feito de que as cualificacións desta convocatoria sexan máis baixas (*vid.* §2.1). Así, a media de formas con códigos non estándar en xuño é de 22,07 e en setembro de 25,82.

A desviación estándar no conxunto dos textos (sen ter en conta as anotacións en *stand-off*, que son computadas á parte en *TEITOK*) é de 10,43 e o rango sitúase entre 84 formas anotadas a nivel de *token* nun texto e 3 anotacións en *stand-off* (así pois, un total de 87) e 3 formas en 3 redaccións, de xeito que ningún dos textos está libre de anotacións deste tipo (en xuño o rango vai de 77 nun texto máis 3 anotacións en *stand-off*, isto é, un total de 80, a 3 formas anotadas en 3 textos, mentres que en setembro vai de 87 a 7, nun texto en cada caso). Na Gráfica 6 observamos a distribución das formas non estándar entre os textos, deixando de novo a un lado as anotacións *stand-off*. Podemos ver como as cifras máis recorrentes se sitúan entre as 10 e as 30 formas non estándar (as cifras que máis se repiten son 12 e 16, en 53 e 47 textos respectivamente). Así mesmo, comprobamos que son escasos tanto os textos con poucos *tokens* anotados (tan só 15 textos con 5 formas ou menos) coma aqueles con 50 ou máis (tamén 15 textos).

Gráfica 6. Distribución das formas non estándar nos textos de *CORTEGAL*



Por outra parte, a distribución dos códigos de formas non estándar entre os diferentes niveis (ortográfico, morfolóxico, léxico, gramatical, semántico e discursivo) é a que se ofrece na Táboa 8[13].

Táboa 8. Códigos descritivos das formas non estándar por nivel

Nivel	**Número de códigos**	**Porcentaxe**
Discursivo	7366	31,79%
Ortográfico	5648	24,38%
Gramatical	3958	17,08%
Léxico	3049	13,16%
Semántico	2004	8,65%
Morfolóxico	1144	4,94%
Total	23.169	

Tal e como se pode observar na Táboa 8, as etiquetas máis frecuentes son as do ámbito discursivo, seguidas polas do nivel ortográfico. Unhas e outras supoñen máis do 50% das etiquetas asignadas. As menos frecuentes son as do ámbito morfolóxico, relativas á flexión, cun 5% dos códigos atribuídos.

[13] Uns poucos códigos do ámbito léxico e do ortográfico poden aparecer máis dunha vez asociados ao mesmo *token* (por exemplo, O_cons_su, cando se produce unha selección inadecuada de dúas consoantes, como en *susesibamente*, ou dous códigos L_w_su cando un calco parcial dunha unidade complexa presenta unha transferencia directa do español no seu interior, como ocorre con *perrito quente*). Neste caso, *TEITOK* só computa un código, co cal estas poucas etiquetas duplicadas foron computadas manualmente.

A preponderancia dos códigos no nivel discursivo explícase pola existencia de moitos problemas relativos ao uso dos signos de puntuación, cun total de 5979 etiquetas, dos cales máis do 60% corresponden a omisións, particularmente de comas. Na Táboa 9 ofrécense os 10 códigos con máis de 500 asignacións, onde se pode comprobar como o código que identifica estas omisións de signos de puntuación, D_pm_om, é o máis frecuente (15,84% do total de etiquetas do corpus). Na táboa tamén podemos ver como entre as etiquetas máis frecuentes existen códigos pertencentes aos seis niveis discursivos:

Táboa 9. Códigos descritivos con máis de 500 ocorrencias

Código	**Valor**	**Número de códigos asignados**
D_pm_om	Omisión dun signo de puntuación	3670
L_w_su	Selección dunha unidade léxica non estándar con diferenzas que van máis alá do xénero ou da acentuación	2887
O_ac_om	Omisión de acento gráfico	2363
S_w_su	Selección dunha unidade léxica inadecuada desde o punto de vista semántico ou combinatorio	1391
D_pm_su	Selección inadecuada dun signo de puntuación	1223
D_pm_ad	Adición dun signo de puntuación	919
G_num_su	Selección de número non estándar	854
O_cons_ad	Adición de consoante	798
O_ac_ad	Adición de acento gráfico	672
M_v_su	Selección dunha forma flexiva verbal non estándar	520

Así pois, as desviacións máis recorrentes son as que afectan ao emprego dos signos de puntuación, os ortográficos relativos á acentuación (*vid.* o capítulo de López Sández e Lorenzo Herrera neste volume), á selección do léxico (sobre todo pola presenza de moitas transferencias do español, *vid.* o capítulo de Álvarez de la Granja nesta obra) e ao seu emprego contextual adecuado, á selección do número (esencialmente por concordancias inadecuadas, *vid.* o capítulo de Cidrás neste volume) e á flexión verbal (normalmente, tamén por influencia do español, *vid.* Álvarez de la Granja 2020b). Ademais, existe un considerable número de formas en que se produce a adición dunha consoante (O_cons_ad), pero debe sinalarse que un 85% das palabras etiquetadas con este código son as voces *producto*, *producción* e outras formas da

mesma familia de palabras (*productor*, *productividade*...), moi recorrentes no corpus dado que un dos comentarios trata sobre o consumo e a produción.

Un segundo nivel de anotación atende á orixe da desviación do estándar, pero como xa indicamos en §2.5.4, o emprego destes códigos está restrinxido a uns poucos casos predeterminados, de xeito que non todas as formas non estándar reciben unha etiqueta deste tipo. Ademais, unha mesma desviación pode ser explicada con máis dun código. En consecuencia, calquera análise e comparativa que se realice está condicionada por estas características. Na Táboa 10[14] ofrécese información sobre a distribución dos códigos asignados, que son un total de 7995[15] e se presentan ordenados de maior a menor frecuencia.

Táboa 10. Distribución dos códigos de orixe

Orixe	**Códigos**	**Número de códigos asignados**	**Porcentaxe**
sp	O_sp M_sp G_sp L_sp S_sp	5033	62,95%
gal	M_gal L_gal G_gal S_gal	820	10,26%
spadapt	M_spadapt L_spadapt	649	8,12%
anal	O_anal M_anal L_anal	592	7,40%
hc	L_hc G_hc	576	7,20%
or	G_or	206	2,58%
for	O_for M_for L_for S_for	109	1,36%
mix	L_mix	10	0,13%
		7995	

[14] Na Táboa 4 poden verse os valores e os casos en que se emprega cada elemento.

[15] Tamén desta volta computamos manualmente aqueles casos en que unha mesma etiqueta aparece por duplicado asociada a un *token*, como ocorre, por exemplo, cun caso de colocación enclítica en vez de proclítica dun pronome átono e ao mesmo tempo de cheísmo. Neste caso, o clítico etiquétase con dous códigos G_hc.

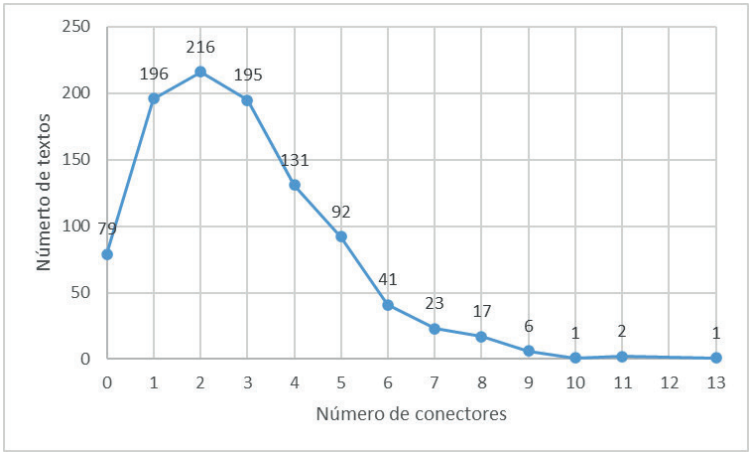
Tal e como se pode comprobar na Táboa 10, as etiquetas máis recorrentes, con moita diferenza sobre as demais, son as que atribúen a orixe da forma non estándar a unha transferencia do español. Precisamente, o código máis frecuente é L_sp, do que hai 2270 exemplos, e que identifica as unidades léxicas transferidas directamente desde o castelán. A segunda orixe máis frecuente é aquela que identifica as formas procedentes de variedades non estándar do galego, sendo M_gal o código máis repetido neste grupo, con 621 resultados. A razón estriba sobre todo na frecuencia no corpus de formas verbais que se poden explicar por transferencia desde o galego popular e tamén pola presenza de numerosos exemplos do demostrativo neutro con <e> (*esto*, *eso* e en menor medida *aquelo*). Nótese, en calquera caso, que estas últimas formas e unha boa parte das primeiras tamén se poden explicar desde o español e reciben igualmente os correspondentes códigos de orixe. Superan tamén as 500 ocorrencias os casos en que se produce a adaptación dunha unidade léxica ou dun elemento flexivo desde o español, así como as formas que se poden explicar como hipercorreccións e como analoxías.

3.4. Os conectores que vinculan enunciados

Finalmente, ofrecemos os datos cuantitativos sobre os conectores empregados no corpus para vincular enunciados. Son 2839 os códigos deste tipo asignados (o que supón unha media de 2,84 conectores deste tipo por texto). A media en xuño é de 2,93 e en setembro de 2,04, cunha diferenza de case un conector entre unha e outra convocatoria.

O rango sitúase entre 13 conectores nunha redacción (da oportunidade de xuño) e ningún en 79 textos (o 16,67% dos textos de setembro non teñen ningún conector entre enunciados e o 6,90% dos de xuño), mentres que a desviación estándar é de 1,93. Na Gráfica 7 obsérvase a distribución dos conectores entre os 1000 textos. Nela, podemos comprobar como a cantidade máis repetida é 2 (216 textos teñen dous conectores deste tipo) e, a medida que se incrementa a cifra de conectores, diminúe o número de textos.

GRÁFICA 7. Distribución dos conectores nos textos de *CORTEGAL*



A distribución por tipo é a que se recolle na Táboa 11[16], onde a ordenación atende á frecuencia de cada tipo de conector.

TÁBOA 11. Distribución dos códigos dos conectores que vinculan enunciados

Código	**Número de códigos**	**Porcentaxe**
ad	913	32,16%
contraarg	616	21,70%
cons	389	13,70%
concl	356	12,54%
ord_sp	225	7,93%
exempl	100	3,52%
ord_t	98	3,45%
caus	57	2,01%
reform	33	1,16%
coment	29	1,02%
confirm	13	0,46%
disjunc	7	0,25%
fin	3	0,11%
2839		

[16] Na Táboa 5 pode atoparse o valor de cada código.

Tal e como se pode comprobar na Táboa 11, os conectores máis frecuentes son os de adición, seguidos dos contraargumentativos, dos consecutivos e dos conclusivos. Non resultan estraños estes resultados, tendo en conta o carácter argumentativo dos textos. O conector máis frecuente é *tamén*, etiquetado en 421 casos. Os outros conectores que superan os 100 códigos son *pero*, *en conclusión*, *e* e *ademais*. Lémbrese en todo caso, que estes elementos só se etiquetan cando serven para unir enunciados.

4. Conclusión

Na primeira parte deste traballo presentamos os criterios metodolóxicos empregados tanto na selección da mostra dos textos que conforman *CORTEGAL* como na transcrición e na anotación das redaccións. Por razóns de espazo, a presentación sobre estas dúas últimas tarefas limítase a aspectos xerais, pero na páxina do corpus poden atoparse diferentes documentos que complementan a información proporcionada aquí.

Ao longo desta primeira parte, tamén demos conta das posibilidades de visualización, tanto dos textos como das súas anotacións, que ofrece *TEITOK*, a plataforma en que se desenvolve o corpus. O feito de poder visualizar os textos nas súas diferentes capas de estandarización, así como a identificación mediante distintas cores das formas non estándar, contribúen a que o corpus poida ser empregado nas aulas para traballar diferentes aspectos lingüísticos.

A anotación das formas non estándar realizada no corpus sitúase no marco metodolóxico da denominada «Análise informatizada de erros» (ou «Análise de erros asistida por ordenador»), pero o deseño do sistema de anotación, para o que non existe un modelo estandarizado, é específico de *CORTEGAL* e responde aos obxectivos do corpus e ás especificidades dos textos que o conforman. Optamos por unha análise bastante detallada das formas non estándar, cun sistema multicapa organizado en seis niveis lingüísticos e xerarquizado desde os niveis onde se levan a cabo estandarizacións vinculadas á forma dos *tokens* ata aqueloutros en que é preciso atender a aspectos contextuais para realizar a normalización. Ademais de codificar o tipo de desviación do estándar (nivel e aspecto afectado, e diferenza entre a forma estándar e a escrita polo alumnado) consideramos de interese informar nalgúns casos, e nun campo de anotación distinto, da orixe das formas diverxentes, así como ofrecer unha primeira aproximación á anotación doutros aspectos discursivos, máis alá das desviacións do estándar, mediante a identificación e codificación dos conectores entre enunciados.

Na segunda parte do traballo presentamos algúns dos resultados atopados en *CORTEGAL*. Limitámonos a aspectos xerais, posto que a análise detallada das distintas cuestións se leva a cabo nos restantes traballos desta mesma obra, en traballos previos (como Álvarez de la Granja, 2020a e Álvarez de la Granja, 2020b) así como, agardamos, en futuras publicacións.

De acordo coa análise presentada, un texto que atenda ás medias de *CORTEGAL* ten ao redor de 225 palabras (113 diferentes), 9 enunciados de 27 palabras cada un, e entre 4 e 5 parágrafos con 2 ou 3 enunciados. Para unir os diferentes enunciados, emprega tan só 3 conectores. Por outro lado, o texto ten 22 ou 23 formas ou secuencias non estándar, unha cifra que parece bastante elevada tendo en conta a extensión das redaccións. En calquera caso, a presentación realizada responde sobre todo ás características dos textos da convocatoria de xuño, moito máis próximos á media dado que supoñen un 89,8% do total das redaccións do corpus. Os textos de setembro conteñen menos palabras, menos lemas ou palabras diferentes e menos enunciados, pero os que inclúen son máis longos e están vinculados con menos conectores. Ademais, un texto desta convocatoria ten 4 formas non estándar máis ca un texto da convocatoria de xuño. Estas diferenzas van ligadas ás que atopamos nas cualificacións, posto que os textos da primeira convocatoria teñen unha nota media considerablemente máis alta ca os da segunda, cunha diferenza que supera os 1,7 puntos.

As desviacións do estándar son sobre todo relativas ao emprego dos signos de puntuación, á acentuación e á selección de léxico (léxico non normativo ou inadecuado ao contexto). Hai tamén algúns problemas relacionados coa concordancia de número, coa ortografía e coa conxugación verbal. Unha boa parte destas desviacións do estándar responden á transferencia ou influencia do castelán. Precisamente, o capítulo de López-Sández e Lorenzo- Herrera deste traballo, que aborda a acentuación, e o capítulo de Álvarez de la Granja, que trata o léxico non estándar, afondan sobre esta cuestión e amosan ben ás claras a relevancia cuantitativa da influencia do español nas desviacións da norma.

Non queremos rematar este traballo sen poñer de manifesto que a complexa anotación levada a cabo en *CORTEGAL* é susceptible, evidentemente, de correccións e ampliacións. Por un lado, tal e como indicamos en §2.5.6, non se nos escapa a existencia de posibles erros e incoherencias na anotación realizada, eivas e inconsistencias que iremos corrixindo a medida que sexan detectadas. Por tal motivo, as cifras que ofrecemos aquí poden non coincidir exactamente cos resultados que proporcione o corpus en futuras buscas,

aínda que, en calquera caso, os cambios non serán significativos. Por outro lado, consideramos de interese ampliar a anotación e refinala, cando cumpra, en futuras versións do corpus, tanto no relativo ás formas non estándar como a outros aspectos que nos axuden a coñecer as características da escrita en lingua galega do estudantado de Galicia ao remate da educación secundaria, particularmente en relación cos aspectos discursivos.

Tanto o corpus, co seu sistema de buscas, como os documentos empregados ao longo da súa elaboración están dispoñibles en aberto para calquera persoa, sexa persoal investigador, profesorado de lingua galega ou calquera outra/o usuaria/o interesada/o. Así mesmo, as persoas responsables do corpus estamos abertas ás súas indicacións e suxestións tanto para correccións de erros como para melloras e ampliacións de máis longo alcance.

O noso obxectivo cando comezamos coa elaboración de *CORTEGAL* en 2017 era deseñar unha ferramenta que servise para coñecer, con datos obxectivos e representativos, como escribe o noso alumnado e, particularmente, aínda que non só, para detectar as principais eivas e carencias na destreza da escritura en lingua galega estándar, pois partiamos da hipótese, demostrada en traballos previos (Silva Valdivia, 2010; López Meirama e Álvarez de la Granja, 2014) e corroborada posteriormente (Loredo e Silva Valdivia, 2020), de que existían algunhas deficiencias importantes en distintos niveis lingüísticos. A análise dos datos confirma unha vez máis esta hipótese, pero *CORTEGAL*, ademais, permite manexar de primeira man os textos que sustentan a súa veracidade. Agardamos que este recurso contribúa tanto a tomar conciencia da necesidade de adoptar medidas para mellorar a destreza da escritura na variedade estándar da lingua galega como a participar directamente nesa mellora mediante a súa utilización directa nas aulas.

Referencias bibliográficas

Abel, A., Glaznieks, A. e Culy, C. (2012). *KoKo German L1 Learner Corpus v2*, Eurac Research CLARIN Centre. http://hdl.handle.net/20.500.12124/11

Abel, A., Glaznieks, A, Nicolas, L. e Stemle, E. (2014). KoKo: an L1 Learner Corpus for German. En N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk e S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 2414-2421). European Languages Resources Association.

Abel, A., Glaznieks, A., Nicolas, L. e Stemle, E. (2016). An extended version of the KoKo German L1 Learner corpus. En A. Corazza, S. Montemagni e G. Semeraro (Dirs.), *Proceedings of the Third Italian Conference on*

Computational Linguistics CLiC-it 2016 (pp. 13-18). Accademia University Press. https://doi.org/10.4000/books.aaccademia.1743

Alonso-Ramos, M. (2016). Spanish learner corpus research. Achievements and challenges. En M. Alonso-Ramos (Ed.), *Spanish Learner Corpus Research Current trends and future perspectives* (pp. 3-31). John Benjamins. https://doi.org/10.1075/scl.78.01alo

Álvarez de la Granja, M. (2018). Corpus de textos de estudantes galegos (CORTEGAL). Aspectos metodolóxicos. En M. Díaz, G. Vaamonde, A. Varela, M. C. Cabeza, J. M. García-Miguel e F. Ramallo (Eds.), *Actas do XIII Congreso Internacional de Lingüística Xeral* (pp. 55-62). Universidade de Vigo. http://cilx2018.uvigo.gal/actas/resumos/655842.html

Álvarez de la Granja, M. (2020a). Análise da conxugación verbal no *Corpus de textos galegos escritos por estudantes no ámbito académico* (CORTEGAL). En L. de Castro Moutinho, R. L. Coimbra e A. Gómez Bautista (Coords.), *Línguas Minoritárias e Variação Linguística* (pp. 150-173). Universidade de Aveiro. https://doi.org/10.34624/rj68-vz44

Álvarez de la Granja, M. (2020b). O infinitivo flexionado no *Corpus de textos galegos escritos por estudantes no ámbito académico* (CORTEGAL). *Madrygal. Revista de estudios gallegos*, 23, 15-38. https://doi.org/10.5209/madr.73064

Álvarez de la Granja, M. e López Meirama, B. (2021). La presencia del español en el léxico disponible del gallego. El centro de interés el cuerpo humano. En M. Serrano Zapata e M. Á. Calero Fernández (Eds.), *Aplicaciones de la disponibilidad léxica* (pp. 115-145). Tirant Humanidades.

Amaro, R., Correia, S., Gramacho, C. e Mendes, A. (2020). Automatização no diagnóstico de nível de língua: anotação e versatilidade dos recursos para PLE. *Revista da Associação Portuguesa de Linguística*, 7, 1-20. https://doi.org/10.26334/2183-9077/rapln7ano2020a1

Burnard, L. (2005). Metadata for corpus work. En M. Wynne (Ed.), *Developing linguistic corpora: a guide to good practice*. Oxbow Books. https://users.ox.ac.uk/~martinw/dlc/chapter3.htm

Capsada Blanch, R. e Torruella Casañas, J. (2017). Métodos para medir la riqueza léxica de los textos. Revisión y propuesta. *Verba*, 44, 47-408. https://doi.org/10.15304/verba.44.3155

Centre for English Corpus Linguistics (2022). *Learner Corpora around the World*. Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Corder, S. P. (1974). Error analysis. En J. P. Allen e S. P. Corder (Eds.), *The Edinburgh Course in Applied Linguistics* (pp. 122-154, vol. 3). Oxford University Press.

Dagneaux, E., Denness, S. e Granger, S. (1998). Computer-aided error analysis. *System*, 26, 126-174. https://doi.org/10.1016/S0346-251X(98)00001-3

Díaz-Negrillo, A. e Fernández Domínguez, J. (2006). Error Tagging Systems for Learner Corpora. *Revista española de lingüística aplicada*, 19, 83-102.

Díez-Bedmar, M. B. (2021). Error Analysis. En N. Tracy-Ventura e M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 99-104). Routledge. https://doi.org/10.4324/9781351137904-9

Dulay, H., Burt, M. e Krashen, S. (1982). *Language two*. Oxford University Press.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford University Press.

Figueras, C. (1999). La semántica procedimental de la puntuación. *Espéculo. Revista de estudios literarios*, 12. https://webs.ucm.es/info/especulo/numero12/puntuac.html

Gilquin, G. e Granger, S. (2022). Using data-driven learning in language teaching. En A. O'Keeffe e M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (2.ª ed., pp. 430-442). Routledge. https://doi.org/10.4324/9780367076399

Gómez Guinovart, X. e Solla, M. (2017-2022). *DContado*. Universidade de Vigo. https://ilg.usc.gal/dcontado/index.php?lang=gl

González Álvarez, E. (1999). Análisis de los errores léxico-semánticos. En L. Iglesias Rábade (Coord.), *Análisis de los errores del examen de inglés en las pruebas de acceso a la Universidad en el distrito universitario de Galicia* (pp. 207-270). Universidade de Santiago de Compostela.

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal*, 20(3), 465-480. https://doi.org/10.1558/cj.v20i3.465-480

Granger, S. (2017). Learner Corpora in Foreign Language Education. En S. Thorne e S. May (Eds.), *Language, Education and Technology. Encyclopedia of Language and Education* (pp. 427-440). Springer. https://doi.org/10.1007/978-3-319-02237-6_33

Granger, S., Gilquin, G. e Meunier, F. (2015). Introduction: Learner corpus research - Past, present and future. En S. Granger, G. Gilquin e F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 1-5). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.001

Instituto Galego de Estatística (2019). *Enquisa estrutural a fogares. Coñecemento e uso do galego*. Instituto Galego de Estatística. https://www.ige.eu/web/mostrar_actividade_estatistica.jsp?idioma=gl&codigo=0206004

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. En N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk e S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4037-4043). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf

Leech, G. e Wilson, A. (1996). *Recommendations for the Morphosyntactic Annotation of Corpora*. Expert Advisory Group on Language Engineering Standards. https://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/annotate.pdf

Lei 3/1983 de 15 de xuño de Normalización lingüística. 14 de xullo de 1983. *Diario Oficial de Galicia*, nº 84.

López Meirama, B. e Álvarez de la Granja, M. (2014). *Léxico dispoñible do galego*. Universidade de Santiago de Compostela. https://www.usc.gal/libros/es/lingstica/125-lexico-disponible-do-galego-lexico-disponible-do-galego.html

Loredo, X. e Silva Valdivia, B. (Coords.) (2020). *Avaliación da competencia bilingüe nos idiomas galego e castelán do alumnado de 4º da ESO*. Real Academia Galega. https://publicacions.academia.gal/index.php/rag/catalog/view/372/366/383

Lüdeling, A., Adolphs, P., Kroymann, E. e Walter, M. (2005). Multi-level error annotation in learner corpora, *Proceedings from the Corpus Linguistics Conference Series* 1(1). University of Birmingham. http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx

Lüdeling, A. e Hirschmann, H. (2015). Error annotation systems. En S. Granger, G. Gilquin e F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135-158). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.007

McEnery, T. e Richard, X. (2010) What corpora can offer in language teaching and learning. En E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 364-380, vol. 2). Routledge.

MERLIN project (2014). *Annotation guidelines*. https://merlin-platform.eu/docs/Annotation%20guidelines.pdf

Mendes, A., Antunes, S., Janssen, M. e Gonçalves, A. (2016). The COPLE2 corpus: A learner corpus for Portuguese. En N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk e S. Piperidis (Eds.), *Proceedings of the Tenth Language Resources and Evaluation Conference (LREC 2016)* (pp. 3207-3214). European Language Resources Association (ELRA). https://aclanthology.org/L16-1511.pdf

Mikelić Preradović, N. (2020). Označavanje pogrešaka u CroLTeC-u (računalnom učeničkom korpusu hrvatskog kao stranog jezika). *Rasprave: Časopis Instituta za Hrvatski Jezik i Jezikoslovlje*, 46(2), 899-920. https://doi.org/10.31724/rihjj.46.2.24

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H. e Andreas, T. (2012). *Das Falko-Handbuch: Korpusaufbau und Annotationen. Version 2.01*. Institut fur deutsche Sprache und Linguistik, Humboldt-Universitat zu. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/view

Reznicek, M., Lüdeling, A. e Hirschmann, H. (2013). Competing target hypotheses in the Falko corpus. A flexible multi-layer corpus architecture. En A. Díaz-Negrillo, N. Ballier e P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 101-124). John Benjamins. https://doi.org/10.1075/scl.59.07rez

Río, I. del e Mendes, A. (2018). Error annotation in the COPLE2 corpus. *Revista da Associação Portuguesa de Linguística*, 4, 225-239. https://doi.org/10.26334/2183-9077/rapln4ano2018a42

Römer, U. (2011). Corpus Research Applications in Second Language Teaching. *Annual Review of Applied Linguistics*, 31, 205-225. https://doi.org/10.1017/S0267190511000055

Roselló Verdeguer, J. (2015). El texto y sus propiedades: algunas consideraciones de carácter práctico. *Tonos Digital. Revista de estudios filológicos*, 28. https://www.um.es/tonosdigital/znum28/secciones/tintero-7--rosello_texto.htm

Rosen, A. (2015). *CzeSL-MAN - a corpus of non-native speakers' Czech with manual annotation*. Informe técnico. Charles University in Prague. http://utkl.ff.cuni.cz/~rosen/public/2015-czesl-man-en.pdf

Stemle, E. W., Boyd, A., Janssen, M., Lindström Tiedemann, T., Mikelić Preradović, N., Rosen, A., Rosén, D. e Volodina, E. (2019). Working together towards an ideal infrastructure for language learner corpora. En A. Abel, A. Glaznieks, V. Lyding e L. Nicolas (Eds.), *Widening the Scope of Learner Corpus Research. Selected papers from the fourth Learner Corpus Research Conference* (pp. 427-468). Presses universitaires de Louvain.

Silva Valdivia, B. (Dir.). (2010) *Avaliación da competencia do alumnado de 4º da ESO nos idiomas galego e castelán*. Universidade de Santiago de Compostela.

Tenfjord, K., Meure, P. e Hofland, K. (2006). The ASK corpus: A language learner corpus of Norwegian as a second language. En N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk e D. Tapias (Eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1821-1824). European Languages Resources Association. http://www.lrec-conf.org/proceedings/lrec2006/pdf/573_pdf.pdf

Timmis, I. (2015). *Corpus Linguistics for ELT: Research and Practice*. Routledge. https://doi.org/10.4324/9781315715537

Vázquez Rozas, V. e Blanco, M. (2022). Corpus y enseñanza del español. En G. Parodi, P. Cantos-Gómez e C. Howe (Eds.), *Lingüística de corpus en español. The Routledge Handbook of Spanish Corpus Linguistics* (pp. 342-356). Routledge. https://doi.org/10.4324/9780429329296-26