Os problemas léxicos e semánticos en CORTEGAL, *Corpus de textos galegos escritos por estudantes no ámbito académico*

MARÍA ÁLVAREZ DE LA GRANJA[1]
Universidade de Santiago de Compostela, Santiago de Compostela, España[2]

Resumo: CORTEGAL, *Corpus de textos galegos escritos por estudantes no ámbito académico*, é un corpus conformado por 1000 textos extraídos dos exames de Lingua e Literatura galegas das probas de acceso á universidade, que inclúe corrección e anotación das formas non estándares en seis niveis lingüísticos. O obxectivo deste traballo é xustificar a necesidade de establecer un nivel semántico diferenciado do nivel léxico á hora de levar a cabo o proceso de corrección dos textos. Tras presentar as características básicas do corpus, amosaremos o tratamento outorgado aos problemas léxicos e semánticos nunha selección de seis corpus de aprendentes con corrección multinivel, mostrando que o habitual é non establecer niveis diferenciados para os problemas léxicos e para os semánticos. En CORTEGAL, pola contra, propóñense dúas dimensións distintas, o que se xustifica polo feito de que unha mesma palabra pode acumular erros dos dous tipos e requirir correccións diferenciadas.

Palabras Chave: corpus de aprendentes, análise de erros asistida por ordenador, semántica, léxico, galego

1. Introdución

O obxectivo principal deste traballo é xustificar a decisión adoptada en CORTEGAL, *Corpus de textos galegos escritos por estudantes no ámbito académico*, de establecer un nivel semántico diferenciado do nivel léxico á hora de levar a cabo a corrección e anotación dos textos que o conforman. CORTEGAL é un corpus electrónico elaborado no Instituto da Lingua Galega da Universidade de Santiago de Compostela, conformado por textos extraídos dos exames de Lingua e Literatura galegas da Proba de avaliación de bacharelato para o acceso á universidade (ABAU). Todas as formas, usos, ou estruturas non estándares presentes en CORTEGAL están sendo corrixidas e anotadas con etiquetas que, en seis niveis lingüísticos (ortográfico, morfolóxico, léxico, gramatical, semántico e discursivo), sinalan o tipo de desviación con respecto ao estándar encontrada. O noso corpus sitúase, pois, dentro do marco metodolóxico dos estudos

de Análise de erros asistida por ordenador (véxase, por exemplo, Dagneaux, Denness, et al., 1998; Díaz-Negrillo e Fernández Domínguez, 2006; Díez-Bedmar, 2021 ou Stemle, Boyd, et al., 2019).

Estruturamos a nosa exposición de acordo co esquema que se indica a seguir. En primeiro lugar, en § 2 ofrecemos unha breve presentación de CORTEGAL. A continuación, en § 3 facemos un repaso ao tratamento dos problemas léxicos e semánticos nunha selección de seis corpus de aprendentes que permiten corrección de erros en diferentes niveis de estandarización. A seguir, en § 4, explicamos o tratamento asignado aos problemas léxicos e semánticos en CORTEGAL, xustificando a nosa proposta de establecer dous niveis diferenciados. Finalmente, en § 5 ofrecemos unhas conclusións a partir da exposición previa.

2. Cortegal

2.1. Os textos de CORTEGAL

CORTEGAL é un corpus constituído por 1000 textos escritos en galego por alumnado da comunidade autónoma de Galicia. Trátase de textos manuscritos redactados no marco dos exames da materia «Lingua e literatura galegas» da proba de Avaliación de Bacharelato para o acceso á Universidade (ABAU), tradicionalmente coñecida como Selectividade. As probas ABAU son a vía de acceso á universidade maioritaria en España e son realizadas na maior parte dos casos por alumnado que rematou os estudos de Bacharelato no mesmo curso en que fai as probas e que ten normalmente entre 17 e 18 anos.

Concretamente, os textos que conforman o corpus son redaccións elaboradas como resposta á pregunta número 3 dos exames mencionados, correspondentes tanto á convocatoria de xuño como á de setembro do ano 2017 (para máis información sobre aspectos relativos á confección da mostra de CORTEGAL, remitimos a Álvarez de la Granja, 2018). Na pregunta número 3, cada estudante debe elaborar unha redacción de carácter argumentativo, dunha extensión de entre 200 e 250 palabras, en volta de certo tema vinculado cun texto que se ofrece previamente, podendo escoller entre dúas opcións en cada convocatoria. No curso 2016–2017, os temas da convocatoria de xuño estiveron vinculados coa gastronomía e co consumismo e a produtividade, mentres que os da de setembro trataron os conflitos paterno-filiais e as referencias ou modelos da mocidade. O número total de *tokens* de CORTEGAL, incluíndo os signos de puntuación, é de algo máis de 269.000 formas. Se se exclúen tales signos o número total de tokens é de pouco máis de 244.000. Neste cálculo, as

contraccións e as formas con pronomes enclíticos son descompostas nas unidades que as conforman.

Os textos elaborados polo estudantado teñen carácter anónimo e carecemos de datos sociolingüísticos que nos permitan facer unha análise en función de variables deste tipo. Con todo, podemos proporcionar algúns datos que nos ofrecen unha panorámica xeral do vínculo que ten o alumnado de Galicia coa lingua galega: segundo a última enquisa do Instituto Galego de Estatística, realizada en 2018 (Instituto Galego de Estatística, 2019), un 38 % das persoas entre 15 e 29 anos residentes en Galicia ten unicamente o castelán como primeira lingua, un 31 % aprendeu a falar tanto en galego como en castelán e un 27,82 % ten só o galego como lingua inicial (un 3,18 % selecciona outras situacións). Con respecto á lingua en que falan habitualmente, un 31,86 % sinala que en castelán unicamente, un 30,75 % fala máis castelán ca galego, un 18,94 % di empregar o galego sempre e un 18,45 % fala máis galego ca castelán. Así pois, o galego dos textos de CORTEGAL pode ser unha L1 ou unha L2, pero en todo caso debe terse en conta que a materia de lingua e literatura galegas está presente no currículo de Primaria e de Secundaria en todos os cursos e que ademais o galego é lingua vehicular de varias materias ao longo de todo o período formativo do alumnado.[3]

Por outro lado, é certo que o contexto das probas (co establecemento dunha limitación temporal, a necesidade de responder varias preguntas, así como a tensión derivada da importancia destes exames de cara ao acceso á universidade) non contribúe «a crear un clima en el que alumno [sic] pueda realizar la mejor de sus producciones» (González Álvarez, 1999, p. 209). Aínda así, xustamente pola relevancia das probas, parece evidente que o alumnado pon especial empeño en redactar os textos da mellor maneira posible de cara a obter unha boa cualificación e, por tal motivo, consideramos que este tipo de textos son unha boa pedra de toque para coñecer a competencia do alumnado na destreza da escritura na variedade estándar da lingua galega unha vez que remata a educación secundaria e accede á universidade.

2.2 O tratamento dos textos

Os textos de CORTEGAL foron anotados en TEITOK (http://www.teitok.org/), unha plataforma baseada na web para a visualización, creación e edición de corpus que permite a combinación de anotacións textuais e lingüísticas nun único documento en formato TEI/XML (Janssen, 2016) e que se ten empregado na elaboración doutros corpus con anotación informatizada de erros

(en http://www.teitok.org/index.php?action=projects pode encontrarse unha listaxe de corpus que empregan esta plataforma).

Así, por un lado, as redaccións son anotadas con metadatos que ofrecen información cuantitativa sobre elas: número de palabras e lemas, densidade léxica, número de parágrafos e de enunciados, media de palabras por enunciado e de enunciados por parágrafo, así como número de palabras do enunciado máis longo e do máis curto.

Por outro lado, en TEITOK identificamos e etiquetamos os elementos riscados polo alumnado no proceso de redacción dos textos, así como as palabras ou fragmentos engadidos sobre a redacción inicial, de tal modo que os textos son consultables na versión con riscados e na versión final da/do estudante. Así mesmo, asignámoslle a cada forma lema e categoría gramatical e anotamos as formas ou secuencias que se desvían do estándar en seis niveis lingüísticos diferenciados e dispostos na seguinte orde: ortográfico, morfolóxico, léxico, gramatical, semántico e discursivo. Cada texto pode visualizarse en diferentes capas de estandarización correspondentes aos distintos niveis, tendo en conta, en calquera caso, que en cada nivel son visibles as correccións dese nivel e dos previos, destacadas en diferentes cores.

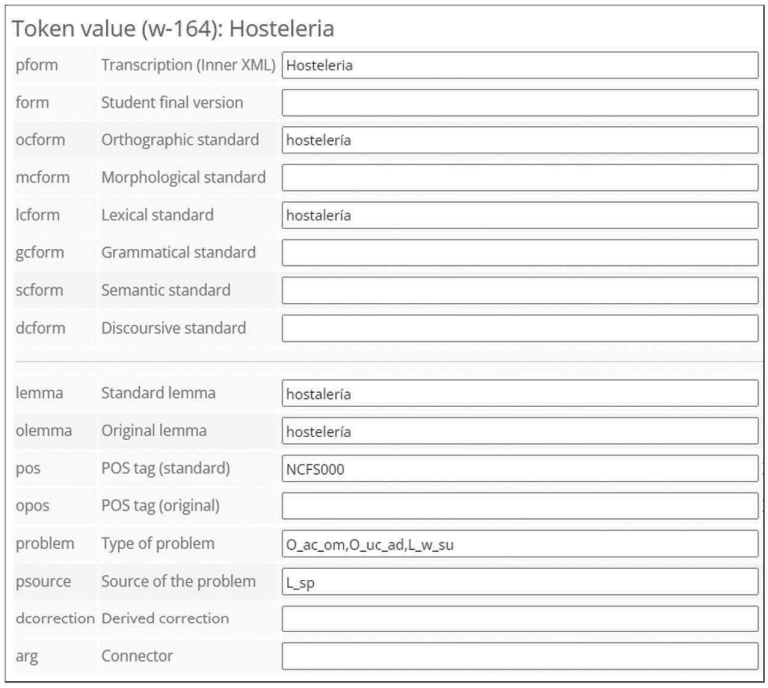
Como sucede en moitos outros corpus de aprendentes,[4] a anotación en CORTEGAL supón dous procesos complementarios: a corrección dos problemas (correspondentes a cada nivel) mediante a asignación de formas normalizadas ou «target hypotheses» e a asignación de códigos que identifican o tipo de problema: «an error can only be annotated if a 'correct' version of the utterance is assumed. Following (Ellis 2009, p. 50) we call this implicit 'correct form' the target hypothesis (TH)» (Reznicek, Lüdeling, et al., 2013, p. 104).

Para levar a cabo a corrección e anotación elaboramos dous manuais con instrucións e exemplos que están accesibles á comunidade científica na páxina do corpus.

Unha mesma palabra pode recibir varios códigos nun mesmo nivel lingüístico, así como códigos pertencentes a diferentes niveis, como no exemplo que se ofrece a seguir (Imaxe 1): na palabra *Hosteleria*, no seguinte contexto,

> a verdade e que a Hosteleria gustalle a xente

acumúlanse 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]. Por outro lado, temos dúas formas normalizadas ou «target hypotheses», unha correspondente ao nivel ortográfico (*hostelería*) e outra ao nivel léxico (*hostalería*).



Imaxe 1. Anotación da palabra *Hostaleria* en CORTEGAL
Fonte: Elaboración propia

Tal e como se pode observar na Imaxe 1, as correccións realízanse para cada nivel lingüístico e, sempre que sexa posible, vanse herdando nos niveis sucesivos, de xeito que, no caso concreto que nos ocupa, a forma correcta *hostalería*, que se ofrece como corrección no nivel léxico, herda as correccións realizadas no nivel ortográfico (engadido de acento e supresión de maiúscula). Por outro lado, en determinados casos ofrecemos tamén unha hipótese explicativa do problema, como nas transferencias doutras linguas: a etiqueta L_sp, que figura no campo «Source of the problem» na Imaxe 1, indica que a diverxencia correspondente ao nivel léxico ten a súa orixe nunha importación desde o español.

Unha vantaxe do emprego de TEITOK para corpus de aprendentes con análise de erros é a posibilidade de establecer, tal e como acabamos de indicar, diferentes capas de estandarización, o que permite que se poidan proporcionar distintas correccións ou «target hypotheses» para unha mesma forma, como

vimos na Imaxe 1. Vexamos un par de exemplos que dan conta da relevancia deste feito.[5]

Nun texto de CORTEGAL rexístrase a forma *articulo*, palabra esdrúxula á que lle falta o acento. Debemos corrixila por *artículo* e asignarlle o código de omisión de acento O_ac_om. Por outra banda, *artículo* é un castelanismo que debemos corrixir pola forma galega estándar *artigo* no nivel léxico, asignándolle o código, xa presentado no apartado previo, L_w_su. Se unicamente puidésemos asignar unha «target hypothesis», que, para obter unha versión estándar plenamente correcta, necesariamente sería *artigo*, estariamos agochando a forma normalizada *artículo*, que é a que xustifica a asignación do código O_ac_om.

Vexamos outro caso cun problema similar. Neste fragmento de CORTEGAL,

> O concurso tiña como tema principal a afirmación de que calquera persoa poda conciñar independentemente dos recursos que dispoña e do tempo

encontramos o emprego de *poda*, forma non normativa do presente de subxuntivo do verbo *poder*, que substituímos no nivel morfolóxico por *poida*, asignándolle o código M_v_su, que se lles atribúe ás flexións verbais non estándares. Pero, ao mesmo tempo, o emprego do subxuntivo nese contexto é inadecuado, de modo que no nivel gramatical corriximos polo indicativo *pode*, asignando o código G_vmt_su, atribuído as formas cunha selección incorrecta de tempo ou modo verbal. Se só fose posible asignar unha forma normativa, que necesariamente tería que ser *pode* para obter unha versión final correcta, estariamos de novo ocultando a forma que xustifica a asignación do código M_v_su, neste caso *poida*.

E, tal e como sinalan Lüdeling e Hirschmann (2015), «an error-annotated corpus which does not provide target hypotheses hides an essencial step of the analysis» (p. 141), pois a codificación de cada forma non estándar está condicionada e xustificada pola forma estándar asignada.

2.3. Os problemas léxicos e semánticos en corpus con anotación multinivel

De acordo co indicado, calquera corpus que asigne formas normalizadas e que permita a asignación destas a diferentes niveis (o cal, segundo o sinalado, parece unha opción moi recomendable, vid. Reznicek, Lüdeling, et al., 2013) debe analizar con coidado cales son as capas ou niveis de estandarización que establece. Por suposto, estas en boa medida están determinadas polo obxectivo do corpus, que pode diferir moito duns a outros. Segundo sinalan Díaz-Negrillo e Fernández Domínguez (2006),

> 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 (p. 89)

Como se pode observar, o ámbito semántico non se menciona nesta revisión xeral. Con todo, este feito non parece responder á desconsideración por parte dos corpus dos usos semanticamente desviados, en que unha palabra se emprega cun significado ou nun contexto que non lle corresponde, senón ao feito de que os problemas deste tipo adoitan ser considerados dentro da dimensión léxica.

Unha ollada a diferentes corpus que permiten asignar «target hypotheses» en varios niveis confirma o afirmado. A selección destes corpus realizouse acudindo a aqueles de entre os considerados en Stemle, Boyd, et al. (2019) que contan con corrección multinivel. Todos eles inclúen producións escritas, aínda que no caso do corpus COPLE2 se recollen tamén algúns textos orais. Na Táboa 1, que se ofrece a seguir, figura a listaxe dos corpus analizados, a lingua ou linguas de cada corpus, as ligazóns aos recursos e os traballos que, canda a análise dos propios corpus, serviron de fonte de información para a súa caracterización.

Táboa 1. Corpus seleccionados

Nome do corpus	**Lingua(s) dos textos**	**Ligazón ao corpus**	**Fonte da información**
COPLE2	L2: portugués	http://teitok.clul.ul.pt/cople2/	Amaro, Correia, et al. (2020); del Río e Mendes (2018)
CroLTeC	L2: croata	http://nlp.ffzg.hr/resources/corpora/croltec/	Mikelić Preradović (2020)
CzeSL-man	L2: checo	http://utkl.ff.cuni.cz/learncorp/	Rosen (2015)
Falko	L2: alemán	https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko	Reznicek, Lüdeling, et al. (2012); Reznicek; Lüdeling, et al. (2013)
KoKo	L1: alemán	https://commul.eurac.edu/annis/koko	Abel e Glaznieks (2017); Abel, Glaznieks, et al. (2016)
MERLIN	L2: checo, alemán e italiano	https://merlin-platform.eu/	Boyd, Hana, et al. (2014); MERLIN project (2014); Wisniewski, Woldt, et al. (2014)

Fonte: Elaboración propia

Algúns destes corpus, concretamente CzeSL-man, Falko e MERLIN permiten proporcionar «target hypotheses» en dous niveis. En trazos xerais, no primeiro nivel lévanse a cabo correccións ortográficas e gramaticais que afectan a elementos illados, mentres que no segundo nivel se realizan outras correccións máis complexas que teñen en conta o contexto: entre elas, as correccións semánticas de palabras existentes pero incorrectamente usadas e as correccións estilísticas (de palabras con marcas diasistemáticas inadecuadas). As palabras «inexistentes» na L2, resultado dunha creación ou dunha transferencia doutra lingua, son corrixidas no primeiro nivel no corpus CzeSL-man, mentres que Falko e MERLIN só corrixen neste nivel as palabras que son transferencia doutras linguas, deixando a normalización doutras formas «inexistentes» para o segundo nivel.

Os restantes corpus analizados permiten a asignación de formas normalizadas en máis niveis, tres en CroLTeC e COPLE2, denominados ortográfico, gramático e léxico, e catro en KoKo, que suma aos anteriores o da puntuación. No nivel léxico considéranse tanto os problemas de selección léxica non natural para unha persoa nativa, como aqueles en que se emprega unha palabra «inexistente» na lingua meta. Ademais, CrolTec engade a omisión de palabras necesarias e a adición de palabras redundantes e KoKo os erros de repetición e redundancia, así como os problemas derivados do emprego de palabras inadecuadas pola súa marcación diasistemática ou diavaliativa.

Observamos, así pois, que por regra xeral os problemas de selección dunha forma inexistente na L2 (que denominaremos léxicos) son corrixidos no mesmo nivel ca os problemas de selección non natural dunha forma, isto é, de atribución a unha palabra dun significado ou uso que non lle corresponde (que denominaremos semánticos). As excepcións son o corpus CzeSL-man, que corrixe os problemas léxicos na primeira capa e os semánticos no segundo, e só parcialmente os corpus Falko e MERLIN, que corrixen na primeira capa as formas que resultan dunha transferencia doutra lingua.

No que respecta aos problemas relativos ás marcas diasistemáticas (por exemplo de rexistro), o corpus KoKo considéraos dentro do nivel léxico, MERLIN e Falko no segundo nivel de corrección e CzeSL-man no primeiro ou no segundo nivel segundo afecten a palabras ou expresións. Os restantes corpus non os consideran.

2.4 Os problemas léxicos e semánticos en CORTEGAL

Tal e como indicamos en § 2.2, a anotación dos textos de CORTEGAL faise en seis niveis lingüísticos: ortográfico, morfolóxico, léxico, gramatical, semántico

e discursivo. Como se observa, consideramos os niveis léxico e semántico como dous niveis diferenciados. Na dimensión léxica corriximos e anotamos aquelas formas ou expresións inexistentes no código normativo galego: entre outras, formas do castelán, como *ahora* ou *basura*, algunhas delas adaptadas ao galego, como *cotidián* ou *chear*,[6] creacións do alumnado como *diferenciativo* ou *moderidade* (por *moderación*) ou construcións que mesturan varias expresións, como *dar lugar a cabo*. Non establecemos, pois, un tipo específico de problema para as formas procedentes doutras linguas, aínda que si damos conta desta circunstancia no campo de anotación «Orixe do problema», mantendo así diferenciada a dimensión descritiva (selección dunha unidade léxica inexistente no código normativo) da explicativa (transferencia desde o español) (Ellis, 1994).

Debe terse en conta, en calquera caso, que o número de transferencias do castelán na lingua galega e particularmente en CORTEGAL (sobre todo na dimensión léxica, pero tamén noutros niveis) é moi alta:[7] temos asignados 2270 códigos L sp (importación directa) e 566 L_spadapt (importación con adaptación) no campo «Orixe do problema». Moitas das formas etiquetadas son formas vivas no galego popular, aínda que non aceptadas no estándar, como *abuelo*, *acostumbrar*, *antoxo*, *calle*…, mentres que outras son formas sen apenas presenza na lingua galega viva, como *aconsexable*, *cosa* ou *chear*. En calquera caso, nun e noutro caso son palabras que se integran plenamente nos textos de CORTEGAL, adaptándose á flexión galega (*sarténs, tutoriais, acarreou, leín*…), e non pezas léxicas situadas entre aspas que mostran ás claras a súa procedencia foránea. E sobre elas realizamos correccións do mesmo tipo que levamos a cabo sobre formas de orixe galega. Así, por exemplo, corriximos a omisión dun acento na última vogal de *conocin*, o emprego de <d> en vez de <z> en *xudgar*, o cambio de <rr> por <r> en *despilfaran* etc.

No nivel semántico, de súa vez, corriximos:

a) a omisión de palabras necesarias para a compleción e comprensión do discurso, como en "non teñen nin as necesidades básicas como ocorre en Venezuela" onde faltaría un adxectivo, por exemplo *cubertas*.
b) a adición de palabras ou expresións innecesarias ou redundantes, como en "É probable que *quizá* a estética dos platos evolucione" onde a palabra *quizá* resulta redundante co adxectivo *probable*.
c) a selección non natural de palabras ou de expresións complexas para expresar determinado contido: os significados ou os usos que a unidade léxica seleccionada pola/polo estudante posúe convencionalmente non se corresponden co significado ou co emprego en CORTEGAL. Así, por exemplo, o uso de *abarcar* en "merecen que *abarquemos* a nosa atención nas súas

> accións ou feitos realizados" suporía un problema semántico, posto que o verbo *abarcar* non posúe o contido «centrar, dirixir» que correspondería ao contexto. A análise dos exemplos anotados mostra como as causas do erro semántico poden ser moi diferentes (calcos semánticos, sobre todo do español; confusión con palabras formalmente próximas; ampliacións semánticas por influencia de sinónimos parciais, lapsos explicables contextualmente; descoñecemento do significado ou uso exacto da palabra ou expresión…).

Pois ben, o establecemento dun nivel léxico diferenciado do nivel semántico é necesario en CORTEGAL na medida en que unha mesma forma pode requirir correccións diferentes para cada unha desas capas, tal e como ocorre nas formas destacadas nos seguintes exemplos:

- os costes de producción son moi *amplios*
- existen varios programas que fomentan que unha persoa *cotidiá* se converta nun cociñeiro de prestixio
- aumento da produción de países como China, provocan un *desvanecemento* da manufacturación e de moitos pequenos comercios

O castelanismo *amplio* debe corrixirse no nivel léxico por *amplo*, pero no nivel semántico cómpre levar a cabo outra corrección, posto que non resulta un adxectivo apropiado ao contexto (fronte a *elevados* ou *grandes*, que si o serían). *Cotidiá*, forma adaptada do castelán *cotidiana*, debe corrixirse pola forma galega estándar *cotiá* no nivel léxico e por *corrente*, por exemplo, no nivel semántico. Finalmente, encontramos de novo unha forma adaptada ao galego dunha voz castelá (*desvanecimiento*), que se corrixe no nivel léxico polo seu equivalente *esvaecemento*. Con todo, esta palabra non é apropiada semanticamente ao contexto, de xeito que levamos a cabo outra corrección no nivel semántico, substituíndoa por *esmorecemento*. Cada unha destas correccións vai acompañada das correspondentes etiquetas identificadoras dos problemas atopados (S_w_su, substitución da palabra adecuada por outra que posúe un significado diferente e non natural no contexto, para o problema semántico, e L_w_su, substitución dunha unidade léxica por outra que non é estándar, para o problema léxico).

Se, tal e como ocorre en varios dos corpus que analizamos no apartado precedente (COPLE2, CroLTeC e KoKo), establecemos un único nivel léxico-semántico, só poderiamos atribuír unha «target hypothesis» (*elevados*, *corrente*, *esmorecemento*), de xeito que agocharíamos a xustificación da asignación do código L_w_su (as formas léxicas estándares *amplos*, *cotiá* e *esvaecemento*).

O tratamento de MERLIN e Falko, que corrixe as formas procedentes doutras linguas no primeiro nivel, mantendo separados estes casos dos problemas semánticos, tampouco nos resulta apropiada por dous motivos. Por un lado, porque a acumulación de problemas léxicos e semánticos non só se encontra nas transferencias, senón que tamén se pode atopar noutras palabras non normativas que non resultan dun proceso de importación, de tal modo que nestes casos o problema se mantén. Así, por exemplo, no seguinte fragmento de CORTEGAL,

> Por iso, se nos non consumimos outros tampouco o estarán facendo, xa que todo vai en cadea; a nosa consumición axuda a que posteriormente alguén poida consumir, e pola contra, a nosa produción tamén axuda a que outra persoa poida producir. Parece todo un xogo de palabras, non si? Pero é a verdade mais *sinceira*.

combínase un problema léxico (L_w_su) consistente no emprego dunha forma non estándar, *sinceira*, creada mediante un fenómeno de analoxía, cun problema semántico (S_w_su), posto que a utilización do citado adxectivo non resulta adecuada nese contexto. Encontrámonos, pois, con dúas formas normalizadas diferentes, respectivamente, *sincera* no nivel léxico e, entre outras posibles opcións, *evidente* no nivel semántico, e debería ser posible ofrecer as dúas «target hypotheses» no sistema de corrección do corpus.

Por outro lado, tal e como xa indicamos, consideramos que o feito de que unha forma resulte dun proceso de transferencia responde a unha clasificación etiolóxica ou explicativa (Ellis, 1994), pero, con independencia da orixe da forma empregada no texto (unha transferencia, unha hipercorrección, unha creación seguindo pautas morfolóxicas produtivas…), o problema é común a formas como *amplio*, *sinceiro* ou *famosidade*: a selección dunha forma léxica inexistente (no sentido de non normativa). En consecuencia, non parece xustificable a corrección en niveis distintos das transferencias e das formas léxicas non estándares con outras orixes.

Este problema non está presente no corpus CzeSL-man, que corrixe no primeiro nivel a selección de calquera forma non estándar, sexa ou non unha transferencia. Con todo, dado que neste primeiro nivel tamén se corrixen os erros ortográficos ou os de flexión, encontrámonos de novo co problema da acumulación de varias «target hypotheses» incompatibles entre si. Así, por exemplo, a forma de CORTEGAL *despilfaran* debe corrixirse no nivel ortográfico por *despilfarran* (posto que o dígrafo <rr> foi substituído por <r>), mentres que no nivel léxico se corrixe por *malgastar*, ao ser *despilfarrar* un castelanismo non aceptado no código normativo. Se a corrección ortográfica e a léxica se realizan no mesmo nivel, só se poderá ofrecer a segunda das formas normalizadas.

No que respecta aos problemas de rexistro, que en varios corpus, tal e como vimos en § 3, se corrixen no nivel léxico ou en todo caso no mesmo nivel que os problemas léxicos e semánticos, en CORTEGAL son corrixidos nun nivel diferente, o discursivo, posto que o problema radica na falta de adecuación a un determinado tipo de texto, de carácter académico e formal. Unha vez máis, poden requirirse correccións múltiples no nivel discursivo e no nivel léxico ou no semántico, o que xustifica o emprego de capas diferenciadas. Así ocorre en

teñen unha pinta asquerosa

onde corriximos no nivel léxico o castelanismo *asquerosa* por *noxenta*, mentres que na dimensión discursiva propoñemos *desagradable*, posto que o emprego de calquera das dúas formas anteriores non parece adecuado nun rexistro académico.

3. Conclusións

A plataforma TEITOK en que se anota CORTEGAL ofrece grandes vantaxes para a construción de corpus de aprendentes porque permite, de maneira doada e de acordo cos intereses das persoas responsables dos corpus, a creación de diferentes capas de corrección independentes entre si:[8] «An error encoding system must therefore allow for the possibility of an in principle arbitrary number of annotation levels» (Lüdeling, Walter, et al., 2005).

Este sistema multinivel permite a asignación de varias formas normalizadas pertencentes a diferentes niveis, dado que unha mesma palabra pode acumular problemas en varios deles que requiran correccións claramente diferenciadas Se asignamos códigos de erro, pero non ofrecemos todas as «target hypotheses» correspondentes, estamos ocultando as formas que precisamente xustifican a anotación, pois, lembremos, «an error can only be annotated if a 'correct' version of the utterance is assumed» (Reznicek, Lüdeling, et al., 2013, p. 104).

Á hora de establecer os diferentes niveis de estandarización, e fronte ao que ocorre nos corpus con corrección multinivel analizados, consideramos necesario manter diferenciados o nivel léxico e o nivel semántico. A razón radica no feito de que nunha mesma forma poden coexistir problemas de selección léxica e de selección semántica que esixen formas normalizadas diferentes, máxime nun corpus, como CORTEGAL, en que o número de problemas de selección de formas non estándares é moi elevado pola alta presenza de transferencias do español. Se unicamente existe un nivel léxico-semántico, ou se os problemas léxicos e os semánticos se corrixen no mesmo nivel, só se poderá ofrecer unha «target hypothesis», que non dará conta das dúas correccións esixidas

pola palabra ou expresión. O mesmo é aplicable aos problemas de rexistro, que poden convivir coa selección dunha palabra inexistente no código normativo (nomeadamente, unha forma doutra lingua) ou dunha palabra semanticamente inadecuada, requirindo a substitución en cada caso por formas estándares distintas. Por tal motivo, parece recomendable establecer un novo nivel (que nós chamamos discursivo) diferenciado dos anteriores.

Debe terse en conta, por outra banda, que a corrección léxica (de formas inexistentes no código normativo) é unha corrección esencialmente de carácter formal e en xeral relativamente doada de aplicar. A corrección semántica ten un carácter máis subxectivo e para realizala cómpre ter en conta o contexto en que aparecen as palabras ou expresións, así como interpretar cal é a intención da/do aprendente. Con respecto á corrección do rexistro, tamén altamente subxectiva, debemos ir máis alá do contexto próximo e ter en conta as características do texto. Trátase, pois, de correccións situadas en ámbitos conceptualmente moi dispares, de xeito que un corpus que ofreza capas de estandarización diferenciadas para cada un deses niveis, sen mesturar as distintas correccións, reflicte máis adecuadamente a diferente natureza destas.

Notas

1 https://orcid.org/0000-0003-4079-2287; maria.alvarez.delagranja@usc.gal

2 Instituto da Lingua Galega.

3 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 alumnado 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).

4 Cremos que CORTEGAL pode considerarse con total propiedade un corpus de aprendentes, malia non ser un corpus de galego lingua estranxeira. Como sinalan os autores do corpus KoKo (Abel, Glaznieks, et al., 2014, p. 2414), poden esgrimirse razóns para utilizar o termo «aprendentes» tamén para os recursos que conteñen textos en L1:

> 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.

O gran número de problemas presentes nos textos do corpus (foron introducidas máis de 23.000 etiquetas identificadoras de desviacións do estándar) apoia sen dúbida esta consideración.

5 Non non estamos referindo aquí á posibilidade de empregar varias «target hypotheses» como diferentes interpretacións alternativas para unha mesma desviación do estándar (vid. Reznicek, Lüdeling, et al., 2013). Esta é unha opción que non empregamos en CORTEGAL.

6 A forma estándar galega é *cotián*, mentres que *cotidián* resulta da adaptación ao galego do español *cotidiano*. No que respecta a *chear* (formado sobre o adxectivo *cheo*), resulta dun calco estrutural do español *llenar*. A forma galega é *encher*.

7 Debe terse en conta que o galego e o castelán son dúas linguas estruturalmente próximas e con moito vocabulario común, o cal, nunha situación de contacto, favorece a transferencia, especialmente da lingua máis prestixiosa á menos valorada. Sobre a presenza de castelanismos no galego, véxase, por exemplo, Dubert García (2005). Véxase, ademais, Álvarez de la Granja e López Meirama (2021) para unha análise das transferencias desde o español no léxico dispoñible do galego, construído a partir das respostas de alumnado de bacharelato.

8 Outras ferramentas que permiten a anotación multicapa son UAM Corpus Tool, EXMARaLDA e ANNIS (Díez-Bedmar, 2021, p. 98).

Agradecementos

Este traballo foi elaborado no marco do proxecto *Corpus de textos gallegos escritos por estudiantes en el ámbito académico. Herramienta para el análisis de la competencia escrita en lengua gallega* financiado por FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación/ Proxecto PGC2018-096069-B-100. O corpus tamén contou coa axuda financeira da Secretaría Xeral de Política Lingüística da Xunta de Galicia a través de convenio coa Universidade de Santiago de Compostela.

Referencias Bibliográficas

Abel, A., Glaznieks, A. Nicolas, L., & 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, & S. Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*

(LREC-2014) (pp. 2414–2421). European Languages Resources Association. http://www.lrec-conf.org/proceedings/lrec2014/pdf/934_Paper.pdf

Abel, A., & Glaznieks, A. (2017). *KoKo: Bildungssprache im Vergleich: korpusunterstützte Analyse der Sprachkompetenz bei Lernenden im deutschen Sprachraum – ein Ergebnisbericht*. Eurac Research 1.0 http://www.korpus-suedtirol.it/KoKo/Documents/Ergebnisse_Dokumentation_gesamt_FINAL.pdf

Abel, A., Glaznieks, A., Nicolas, L., & Stemle, E. (2016). An extended version of the KoKo German L1 Learner corpus. En A. Corazza, S. Montemagni, & G. Semeraro (Dirs.), *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016* (pp. 13–18). Accademia University Press. https://books.openedition.org/aaccademia/1743

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, & 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., & 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, & M. A. Calero Fernández (eds.), *Aplicaciones de la disponibilidad léxica* (pp. 115–145). Tirant Humanidades.

Amaro, R., Correia, S., Gramacho, C., & 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://ojs.apl.pt/index.php/RAPL/article/view/86

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., & Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. En N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)* (pp.1281–1288). European Languages Resources Association. http://www.lrec-conf.org/proceedings/lrec2014/index.html

Dagneaux, E.; Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, 26, 126–174.

Díez-Bedmar, M. B. (2021). Error Analysis. En N. Tracy-Ventura, & M. Paquot (eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 90–104). Routledge.

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

Dubert García, F. (2005). Interferencias del castellano en el gallego popular. *Bulletin of Hispanic Studies*, 83(3), 271–291.

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

González Álvarez, E. (1999). Análisis de los errores léxico-semánticos. En L. Iglesias Rábade (ed.), *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). Instituto de Ciencias da Educación – Universidade de Santiago de Compostela.

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, & S. Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4037–4043). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. En S. Granger, G. Gilquin, & F. Meunier (eds.), *The Cambridge handbook of learner corpus research* (pp. 135–158). Cambridge University Press.

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

MERLIN project (2014). *Annotation guidelines*. https://merlin-platform.eu/docs/Annotation%20guidelines.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., & Hirschmann, H. (2013). Competing target hypotheses in the Falko corpus. A flexible multi-layer corpus architecture. En A. Díaz-Negrillo, N. Ballier, & P. Thompson (eds.), *Automatic Treatment and Analysis of Learner Corpus Data* [Studies in Corpus Linguistics 59] (pp. 101–124). John Benjamins.

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H., & Andreas, T. (2012). *Das Falko-Handbuch: Korpusaufbau und Annotationen. Version 2.01*. Institut fur deutsche Sprache

und Linguistik, Humboldt-Universitat zu Berlin. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/view

Río, I. del, & Mendes, A. (2018). Error annotation in the COPLE2 corpus. *Revista da Associação Portuguesa de Linguística*, 4, 225–239. https://ojs.apl.pt/index.php/RAPL/article/view/42

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., Boyd, A., Janssen, M., Lindström Tiedemann, T., Mikelić Preradović, N. Rosen, A., Rosén, D., & Volodina, E. (2019). Working together towards an ideal infrastructure for language learner corpora. En A. Abel, A. Glaznieks, V. Lyding, & 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.

Wisniewski, K., Woldt, C., Schöne, K., Abel, A., Blaschitz, V., Štindlová, B., & Vodičková, K. (2014). *The MERLIN annotation scheme for the annotation of German, Italian, and Czech learner language*. https://merlin-platform.eu/docs/MERLIN-annotation-scheme.pdf