Corpus de textos de estudantes galegos (CORTEGAL). Aspectos metodolóxicos

María Álvarez de la Granja
Instituto da Lingua Galega-Universidade de Santiago de Compostela
maria.alvarez.delagranja@usc.es

Resumo

O *Corpus de textos de estudantes galegos* (CORTEGAL) é un corpus en período inicial de construción desenvolvido no Instituto da Lingua Galega da Universidade de Santiago de Compostela. Na primeira fase, este corpus estará conformado por 1000 redaccións elaboradas por estudantes de segundo de Bacharelato de Galicia no marco do exame de Lingua e Literatura Galegas da proba de Avaliación do Bacharelato para o acceso á Universidade (ABAU). O obxectivo principal de CORTEGAL é o de contribuír ao coñecemento das principais dificultades que ten o alumnado no momento de elaborar un texto escrito na variedade estándar do galego. Por tal motivo, as formas e secuencias non estándares que figuran nos textos do corpus estarán convenientemente identificadas e anotadas para a súa recuperación a través de diferentes vías. Neste traballo presentamos as características dos textos que conformarán o corpus e os criterios de selección da mostra e amosamos as nosas propostas iniciais en relación coa transcrición e anotación dos textos e coas posibilidades que ofrecerá o sistema de buscas do recurso, que será de acceso libre a través de Internet.

Palabras clave: corpus de aprendentes, ensino e aprendizaxe de linguas, expresión escrita

1. Introdución

O obxectivo deste traballo é presentar algúns aspectos metodolóxicos relativos á construción do *Corpus de textos de estudantes galegos* (CORTEGAL).[1] CORTEGAL é un corpus que se está comezando a desenvolver no Instituto da Lingua Galega da Universidade de Santiago de Compostela e que estará conformado por textos redactados en galego por alumnado de segundo de Bacharelato de Galicia. Na súa primeira fase estará constituído por ao redor de 250.000 palabras.

A finalidade principal deste corpus é a de contribuír a coñecer as principais dificultades que ten o alumnado á hora de escribir na variedade estándar da lingua galega, teña esta como L1 ou como L2. A tal fin, no corpus estarán identificadas as formas que diverxen con respecto ao código normativo do galego e á variedade estándar en que o devandito código se realiza. A análise terá en conta todos aqueles niveis lingüísticos que interveñen na conformación dun texto escrito: ortográfico, morfolóxico, léxico, semántico, sintáctico e textual. Para a marcación no nivel textual teremos en conta as convencións establecidas á hora de redactar un texto nun nivel formal.

No apartado 2 presentaremos as características dos textos e os criterios de conformación da mostra, en §3 abordaremos algunhas cuestións relativas á transcrición e á anotación dos textos, en §4 mostraremos as posibilidades de busca previstas para o recurso, que será de acceso libre en Internet e finalmente no apartado 5 presentaremos algunhas consideracións sobre as utilidades de CORTEGAL.

1 Este traballo desenvólvese dentro do proxecto *Corpus para a análise de dificultades en lingua galega* (2017-CP050), financiado pola Xunta de Galicia a través do convenio asinado entre esta e a Universidade de Santiago de Compostela (Instituto da Lingua Galega) e a través do grupo FILGA (Xunta de Galicia / FEDER, ED431C 2017134).

2. Características dos textos e conformación da mostra

2.1. Os textos

Os textos que conformarán o corpus son textos manuscritos redactados en galego por estudantes de segundo de Bacharelato da comunidade autónoma galega A fonte de tales textos son os exames da materia Lingua e literatura galegas da proba de Avaliación do Bacharelato para o acceso á Universidade (ABAU),[2] máis en concreto as redaccións elaboradas como resposta á pregunta número 3 (valorada con 3 puntos sobre 10), en que se solicita que o/a estudante redacte un texto de ao redor de 200-250 palabras en volta dun tema determinado. Concretamente, os textos seleccionados para esta primeira fase corresponden ás probas de avaliación do curso 2016-2017, tanto da convocatoria de xuño como da de setembro. En cada convocatoria hai dous modelos de exame, de xeito que os textos que incorporaremos no corpus tratarán catro temas diferentes. Estes son os enunciados das catro preguntas número 3, que se vinculan en todos os casos cun texto previo que figura ao comezo do exame e que tamén serve de base para a formulación doutras preguntas da proba:

> 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

Como se pode deducir a partir das preguntas formuladas, os textos que se incorporarán no corpus son textos argumentativos en que o/a estudante debe expresar a súa opinión sobre un tema de actualidade.

Os textos non están identificados co nome do autor ou autora e os únicos datos dos que temos información son a convocatoria a que pertence o exame (xuño ou setembro), a cualificación obtida (tanto na pregunta 3 como na proba completa de lingua e literatura galegas) e a comisión delegada á que corresponde o exame. Existen 26 comisións delegadas e cada unha delas ten asignados unha serie de centros de ensino determinados pertencentes a unha zona xeográfica que normalmente é bastante ampla (a única excepción é a comisión delegada 25, que non ten asignados centros de ensino concretos, senón alumnado con necesidades especiais procedente de calquera parte de Galicia). Por outro lado, os centros e os lugares de procedencia poden ter características moi dispares, como se pode apreciar na Táboa 1, correspondente á Comisión delegada 21, que ofrecemos a modo de exemplo (IES: Instituto de Educación Secundaria; CPR: Centro privado; CIFP: Instituto de Formación Profesional)

[2] Son as probas tradicionalmente coñecidas como de Selectividade.

Centro de ensino	*Localización do centro de ensino*
IES San Tomé de Freixeiro	Vigo (Pontevedra)
IES Castelao	Vigo (Pontevedra)
IES San Paio	Tui (Pontevedra)
IES As Barxas	Moaña (Pontevedra)
IES de Coruxo	Vigo (Pontevedra)
IES A Sangriña	A Guarda (Pontevedra)
IES do Castro	Vigo (Pontevedra)
IES Pedra da Auga	Ponteareas (Pontevedra)
CPR Plurilingüe Montecastelo	Vigo (Pontevedra)
CPR Plurilingüe Compañía de María	Vigo (Pontevedra)
CPR Marcote	Vigo (Pontevedra)
CPR Plurilingüe Lar	Mos (Pontevedra)
CPR San José de Cluny	Vigo (Pontevedra)
CIFP A Granxa	Ponteareas (Pontevedra)

Táboa 1. Centros de ensino asignados á Comisión delegada 21 nas probas ABAU de Galicia (curso 2016-2017)

O feito de traballar con textos extraídos das probas ABAU ten vantaxes e inconvenientes. Os inconvenientes máis relevantes son dous:

- A escasa información dispoñible sobre os/as estudantes. Se os textos se elaborasen especificamente para a conformación do corpus, poderiamos recoller información sobre variables sociodemográficas, sociolingüísticas ou escolares que permitirían análises e cotexos moi relevantes. Por exemplo, sería de gran interese poder contrastar as producións de alumnado que ten o galego como L1 coas do que o ten como L2, para determinar con datos auténticos e representativos as dificultades propias de cada grupo.
- A necesidade de transcribir manualmente os textos (pois traballamos con producións manuscritas), o que supón unha inversión importante de tempo e esforzo.

No que respecta ás vantaxes, estas son probablemente as máis destacadas:

- A dispoñibilidade inmediata dos textos.
- A súa homoxeneidade.
- A garantía (practicamente ao cen por cen) de que o/a estudante elabora a redacción con seriedade e de que utiliza a variedade estándar da maneira que mellor sabe, pois a pregunta forma parte dun exame de acceso ao sistema universitario.
- O valor e representatividade dos textos, posto que as probas de acceso ao ensino universitario teñen unha valoración social engadida que dota o corpus dunha especial relevancia no marco da sociedade galega e poden considerarse unha excelente pedra de toque para determinar o nivel dos estudantes na destreza da expresión escrita ao final da Educación Secundaria.
- A posibilidade de rastrexar o proceso compositivo, dado que tratamos con textos manuscritos que deixan pegada do devandito proceso a través de riscaduras, engadidos... Tal e como se indicará no apartado 2, as opcións de visualización permitirán seguir este proceso, o cal é especialmente relevante para a análise das dificultades que teñen os estudantes e para coñecer os procesos cognitivos subxacentes á composición de textos.

Unha vez sopesadas vantaxes e desvantaxes, optamos por comezar o noso proxecto a partir dos textos das ABAU, tal e como xa indicamos, o que non obsta para que nun futuro se poidan incorporar novos textos procedentes doutras fontes con maior información sobre as características das e dos informantes.

2.2. A mostra

A mostra que utilizaremos para a confección do corpus estará composta, na súa fase inicial, por 1000 textos. Estes textos foron proporcionados pola Comisión Interuniversitaria de Galicia (CIUG) e corresponden tanto á convocatoria de xuño como á de setembro. O número de textos da mostra por comisión delegada e convocatoria será proporcional á cifra total de exames por comisión e convocatoria, coas seguintes salvidades.

- Elimínanse do cómputo os 29 exames correspondentes á Comisión delegada nº 25, que, como xa se sinalou, funciona de maneira diferente ás restantes comisións, pois recolle probas de alumnado con necesidades específicas procedente de toda Galicia e non dunha zona restrinxida xeograficamente.
- Aínda que a distribución de exames entre a convocatoria de xuño e de setembro se sitúa aproximadamente nun 85% e 15% respectivamente, o reparto na mostra será de ao redor do 90% para os exames de xuño e do 10% para os de setembro. Esta circunstancia explícase polo feito de que boa parte do alumnado que realizou o exame na convocatoria de setembro, tamén o fixo previamente na convocatoria de xuño. En consecuencia, se respectamos a súa porcentaxe en setembro, mantendo a distribución inicial (85/15), estaremos outorgándolles máis peso aos informantes que realizan o exame dúas veces. Por este motivo, a partir dos datos fornecidos pola CIUG, determinamos o número de probas ABAU suspensas que houbo para cada comisión delegada en xuño e restamos esa cifra ao número de exames presentados en cada comisión en setembro. As porcentaxes de cada comisión delegada foron calculados a partir das novas cifras, en que se suprimiron os non aptos de xuño, reducindo deste xeito a representatividade dos exames da segunda convocatoria.[3]. O resultado final sitúase, como indicamos, en volta do 90% para xuño e do 10% para setembro.

Por outro lado, e unha vez establecido o número de exames de xuño e o número de exames de setembro correspondentes a cada comisión delegada, a selección das probas farase exclusivamente atendendo á distribución equitativa dos temas (en cada comisión o 50% dos textos corresponderá á opción A e o 50% á opción B, tanto en xuño como en setembro). Se nalgunha comisión non fose posible realizar esta distribución por non existir un número suficiente de exames dunha das opcións, compensarase o desequilibrio na seguinte comisión en que isto sexa posible. Polo demais, a selección dos textos para cada comisión será aleatoria.

3. Transcrición e anotación dos textos

3.1. Transcrición

Está previsto que o corpus se cree, se anote e se consulte a través da plataforma TEITOK (Janssen 2016), que permite diferentes capas de transcrición e de visualización dos textos e que xa se utiliza en corpus de aprendentes como o *Learner Corpus of Portuguese L2*-COPLE2 (Mendes, Antunes, Janssen e Gonçalves 2016). Deste xeito, o noso obxectivo é ofrecer tanto unha transcrición fiel do texto orixinal, incorporando, por exemplo, as riscaduras ou os engadidos feitos polo autor ou pola autora, que serán convenientemente codificados en formato XML, como a versión final proposta por el/ela, de maneira que o usuario poida

[3] A eliminación de "segundos exames" é meramente cuantitativa e non cualitativa, posto que carecemos de datos que nos permitan identificar cales son os exames dos ou das estudantes que suspenderon as ABAU en xuño. A cualificación específica do exame de lingua e literatura galegas non determina o apto/non apto global na proba ABAU.

escoller a visualización que desexe. Nesta visualización, respectaranse os parágrafos establecidos polo ou pola informante, que se identificarán mediante a correspondente codificación en XML. Aínda que noutros corpus, como o COPLE2, tamén se incorporan as correccións ou indicacións da profesora ou do profesor que corrixe o texto, esta opción non se implementará en CORTEGAL.

3.2. Anotación

Os metadatos codificados na cabeceira de cada texto, tamén en formato XML, serán poucos, pois a información que posuímos sobre cada un deles é escasa. Está previsto incorporar os seguintes datos: fonte do texto, ano en que foi escrito, convocatoria (primeira / segunda), lingua do texto, cualificación na pregunta 3 (con catro valores posibles, suspenso, aprobado, notable e sobresaliente), cualificación na proba de lingua e literatura galegas (cos mesmos valores que se acaban de indicar para a pregunta 3), tipo de texto (narrativo, argumentativo etc.), tema do texto e comisión delegada correspondente (o que permitirá asignar unha zona xeográfica ao centro de estudos do informante).[4] Estamos tamén explorando a posibilidade de incorporar información estatística sobre a diversidade léxica e a lexibilidade de cada texto para permitir a súa avaliación e clasificación desde esta perspectiva.

Os textos serán tokenizados, lematizados e anotados gramaticalmente de maneira automática. No caso das formas non estándares, revisaranse o lema e a categoría gramatical asignados automaticamente ou, no caso de que o lematizador non recoñecese a palabra, asignaranse por primeira vez. A proposta de partida é que o lema represente, no seu caso, as diferentes posibilidades flexivas da palabra mantendo as diverxencias fónicas ou gráficas existentes con respecto á forma estándar. Así *luita* e *luitas* (forma estándar *loita*) lematizaranse con *luita*; *aboa*, *abó*, etc. (forma estándar *avó*) con *abó*, *rodilla* e *rodillas* (castelanismo por *xeonllo*) con *rodilla* etc. Deste modo, o lema servirá como un elemento aglutinador daquelas formas que comparten unha mesma diverxencia con respecto ao estándar, o que permitirá recuperar conxuntamente formas como *coibir*, *coibiron*, *coibirán*, *coibisen* etc. (a través do lema *coibir*).

Ademais, levaremos a cabo a clasificación das formas ou secuencias non estándares incluídas nos textos. Para todas elas, ofrecerase unha clasificación bidimensional, que presentamos de forma moi esquemáticas nas seguintes liñas:

- *Dimensión lingüística*. A forma clasifícase nunha categoría que sinala o aspecto ou nivel lingüístico en que se sitúa a diverxencia con respecto ao estándar: puntuación, flexión, concordancia, acentuación gráfica... A nosa intención é identificar diverxencias no nivel ortográfico, morfolóxico, léxico, semántico, sintáctico e textual.
- *Dimensión descritiva*. A forma sitúase nunha categoría que nace da comparación entre ela e a forma estándar correspondente e que reflicte a relación ou a diferenza existente entre ambas desde a perspectiva do texto orixinal: omisión, adición, próclise por énclise...

Ademais, nalgúns casos introducirase unha terceira dimensión:

- *Dimensión explicativa*. A forma sitúase nunha categoría que tenta explicar a razón do seu uso: castelanismo, dialectalismo, confusión, lapsus...

[4] Algún dos datos indicados (por exemplo, a fonte ou o tipo de texto) son comúns a todos os textos nesta primeira fase, pero son introducidos prevendo a posibilidade de ampliar o corpus nun futuro con outros textos que teñan características distintas (por exemplo, outra fonte ou outro tipo de texto).

A modo de exemplo, a ausencia dunha coma nun texto sería clasificada na dimensión lingüística como puntuación e na dimensión descritiva como omisión; a voz *rodilla* sería anotada coas marcas léxico (dimensión lingüística), substitución (dimensión descritiva) e transferencia do castelán (dimensión explicativa) etc.

A tipoloxía de formas non estándares está en fase de elaboración e para a súa construción tomamos como punto de partida sistemas similares empregados en corpus de aprendentes con análise informatizada de erros (vid. por exemplo Dagneaux, Denness e Granger 1998; Díaz-Negrillo e Fernández-Domínguez 2006; Lüdeling e Hirschmann 2015). Cómpre realizar, en calquera caso, as necesarias adaptacións derivadas das peculiaridades da lingua analizada e dos informantes (estudantes que teñen o galego como L1 ou L2, pero non como LE, fronte ao que é habitual nos corpus de aprendentes).[5] Neste sentido, propoñemos xa de partida unha adaptación terminolóxica, utilizando a expresión *forma non estándar* no canto de *erro*, termo moi pouco apropiado para a etiquetaxe de certas voces galegas alleas ao código normativo, como os dialectalismos. Por outro lado, para a conformación da clasificación das formas non estándares acudimos tamén á información que nos proporcionan diversos traballos centrados na análise das dificultades que teñen os estudantes galegos á hora de empregar a variedade estándar desta lingua, dos que poden ser un bo exemplo os dicionarios ou glosarios de dúbidas (por exemplo, Fernández Salgado 2004; Hermida 2004; Feixó, Pena e Rosales 2010). Finalmente, cómpre sinalar que, aínda que a clasificación dunha forma non estándar sempre é subxectiva, a dimensión explicativa está suxeita a un maior grao de subxectividade ca as outras, de xeito que a aplicabilidade e alcance desta dimensión deben ser estudados con especial atención antes de establecer o sistema definitivo. Con todo, non quereriamos renunciar a ela posto que dá conta, entre outros aspectos, dalgunhas das clasificacións tradicionais de formas non estándares en galego.

Por outra banda, e sempre que sexa posible (pode resultar complexo, por exemplo, nalgúns problemas de carácter discursivo), asignaráselle a cada forma non estándar a forma estándar correspondente. Así, por exemplo, a *aboa* asignaráselle *avoa*, a *rodillas* a forma *xeonllos* ou a *Me chamo* a secuencia *Chámome*. Como sinalan Lüdeling e Hirschmann (2015: 141), "an error-annotated corpus which does not provide target hypotheses hides an essencial step of the analysis", pois a clasificación está condicionada por unha determinada interpretación do anotador, que debe establecer previamente cal é a forma estándar correspondente (véxase a caracterización da dimensión descritiva establecida liñas arriba).

Finalmente, esa forma estándar irá acompañada en determinados casos[6] dun lema estándar que permitirá recuperar conxuntamente diferentes formas flexivas que presentan algunha diverxencia con respecto ao estándar (por exemplo, *aboa*, *abó*, *avo* e *aboas* recuperaranse conxuntamente a través do lema estándar *avó*; *rodilla* e *rodillas* a través do lema estándar *xeonllo*).

---

[5] Con todo hai excepcións. Por exemplo, o corpus de alemán KoKo é considerado un corpus de aprendentes malia estar constituído por textos escritos por estudantes que teñen esa lingua como L1. Abel, Glaznieks, Nicolas e Stemle (2014: 2414) xustifican esta consideración:

> 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.

[6] Carece de sentido en exemplos como o de *Me chamo* (forma estándar *Chámome*) previamente presentado.

De acordo co indicado, e a modo de exemplo, a unha forma como *coibiron* asignaráselle o lema *coibir*, a forma estándar *cohibiron* e o lema estándar *cohibir*, á palabra *rodillas* o lema *rodilla*, a forma estándar *xeonllos* e o lema estándar *xeonllo* etc.

Por outro lado, exploraremos a posibilidade de identificar e anotar nos textos algúns elementos que contribúen a dotalos de cohesión textual, como os conectores e as referencias, de tal xeito que o corpus poida ser aproveitado en maior medida por investigadores que traballan no ámbito da gramática textual.

4. Sistema de buscas

O corpus será accesible de forma libre a través de Internet e a nosa intención é que a aplicación de consulta ofreza estas posibilidades de busca e filtraxe:

1. Busca por lema e por forma de palabra. O usuario poderá buscar en todo o corpus a través do lema ou a través da forma gráfica escrita polo ou pola estudante (ou combinando un lema e unha forma gráfica) e poderá restrinxir a busca a unha determinada categoría gramatical do elemento buscado. O resultado poderá visualizarse nunha concordancia en formato KWIC ou nun contexto máis extenso, aínda que en ambos os casos este poderá ser ampliado. A partir da busca realizada, o usuario terá acceso (neste tipo de busca e nas que seguen) a información extraída dos metadatos. Ademais, todas as formas non estándares irán asociadas con información sobre a súa clasificación na tipoloxía de formas non estándares e sobre a forma e lema estándar que lles corresponden.

2. Busca pola clasificación das formas non estándares. Como resultado desta busca, o usuario obterá unha concordancia coas formas ou secuencias non estándares que responden á clase ou clases seleccionadas.

3. Busca polas formas estándares ou polos lemas estándares. O resultado desta busca será unha concordancia coas formas non estándares que teñen asignada a forma estándar introducida na busca ou, no seu caso, o lema estándar.

4. Filtraxe pola información que figura nos metadatos. As buscas poderán restrinxirse aos textos que cumpren determinadas características, como unha determinada temática ou cualificación na proba de lingua e literatura galegas.

Por outra banda, unha vez realizada a consulta, e tal e como xa sinalamos no apartado 2, o usuario poderá visualizar o texto na súa forma final ou na versión que incorpora todos os cambios realizados pola ou polo estudante ata chegar ao texto definitivo, con indicación das riscaduras, engadidos ou correccións realizados no proceso de composición do texto.

5. Consideracións finais

Coidamos que a elaboración do corpus CORTEGAL é relevante para a investigación sobre o galego en varios sentidos. Por un lado, esta ferramenta fornecerá exemplos reais das dificultades que teñen os estudantes de Educación Secundaria á hora de escribiren na variedade estándar da lingua galega, o cal permitirá valorar cualitativa e cuantitativamente esas dificultades.

Por outro lado, o coñecemento que proporcionará o corpus sobre as dificultades que ten o alumnado á hora de construír textos escritos será un punto de partida importante para elaborar recursos de axuda á escritura, que contribúan á mellora da competencia escrita en galego. Os datos extraídos do corpus permitirán tamén enriquecer libros de texto e outros materiais didácticos xa existentes e ilustralos con exemplos reais. O noso corpus tamén será relevante para a lexicografía, ao facilitar a incorporación nas entradas dos dicionarios de información sobre as dificultades máis comúns asociadas con certas palabras.

En terceiro lugar, o corpus permitirá analizar as características xerais da lingua escrita dos estudantes galegos actuais. A este respecto, debe indicarse que non existe para o galego un recurso semellante a este, pois os corpus de lingua escrita que posuímos recollen sobre todo textos literarios, xornalísticos e técnicos. O noso recurso complementará e enriquecerá a información que estes proporcionan sobre o galego escrito.

Finalmente, coidamos que a elaboración deste corpus contribuirá a desenvolver liñas de investigación con pouco percorrido nos estudos sobre o galego, como a gramática textual e a composición de textos, e abrirá novas vías, os corpus de aprendentes e a "análise informatizada de formas non estándares", liñas que, malia a súa mocidade, deron xa abundantes e proveitosos froitos noutras linguas.

Referencias bibliográficas

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)*. Reykjavik: European Languages Resources Association, 2414-2421. <http://www.lrec-conf.org/proceedings/lrec2014/pdf/934_Paper.pdf> [Consultado o 10 de maio de 2018].

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

Díaz-Negrillo, A. e Fernández-Domínguez, J. (2006). "Error tagging systems for learner corpora". *Revista Española de Lingüistica Aplicada*, 19, 83-102.

Feixó, X, Pena, X.R. e Rosales, M. (2010). *Galego século XXI: Nova guía da lingua galega*. (2ª ed.). Vigo: Galaxia.

Fernández Salgado, B., dir. (2004). *Diccionario Galaxia de usos e dificultades da lingua galega*. Vigo: Galaxia.

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*. Cambridge: Cambridge University Press, 1-5.

Hermida, A. (2004). *Consultor Cumio de galego*. (2ª ed.).Vigo: Edicións do Cumio.

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)*. Portorož: European Language Resources Association (ELRA), 4037-4043. <http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf > [Consultado o 10 de maio de 2018].

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*. Cambridge: Cambridge University Press,135-158.

Mendes, A., Antunes, S, Janssen, N. 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 International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož: European Language Resources Association (ELRA), 3207-3214. <http://www.lrec-conf.org/ proceedings/lrec2016/pdf/439_Paper.pdf> [Consultado o 10 de maio de 2018].