

CORTEGAL. Corpus de textos gallegos escritos por estudiantes en el ámbito académico

María Álvarez de la Granja / Xulio Sousa

Instituto da Lingua Galega-Universidade de Santiago de Compostela
maria.alvarez.delagranja@usc.gal / xulio.sousa@usc.gal

III Workshop INTELE

Madrid, 13-14 septiembre 2022

1. Presentación

CORTEGAL es un corpus que ofrece anotación de las formas no estándares conformado por 1000 textos argumentativos de entre 200 y 250 palabras extraídos de los exámenes de lengua gallega de las pruebas ABAU de Galicia (curso 2016-2017). Su objetivo es conocer las características de la producción escrita en lengua gallega del alumnado de Galicia al finalizar la educación secundaria, así como servir como herramienta de ayuda en el aula. El corpus, consultable en <http://ilg.usc.gal/cortegal/es/index.php?>, está en fase de revisión.

tamén supuxo a popularidade da cocaña.
Pero hai programas que nunca van a
pasar de moda como "Arguiñano".

En mi opinión, a cocaña cada vez
vai ser mais popular porque é un
campo moi amplo a novidade e
vánse des cubrir coas novas e a
xente vaille gustar cada vez máis
a gastronomía

2. Transcripción

Los textos fueron transcritos y anotados en TEITOK, una plataforma basada en la web para la visualización, creación y edición de corpus que permite la combinación de anotaciones textuales y lingüísticas en un único documento en formato TEI/XML. Se transcriben y etiquetan las formas tachadas por el/la estudiante y también se utiliza una etiqueta específica para las formas claramente añadidas a posteriori y para las de lectura dudosa.

>Arguiñano</tok><tok id="w-179" lemma=""" pos="Fe"></tok><tok id="w-219" dcform=". " problem="D_pm_om"><ee/></tok></p> <p id="p-4"><tok id="w-180" gcform="Na" problem="G_det_om" psource="G_sp" lemma="en" pos="SP">En</tok> <tok id="w-181" lcform ="miña" olemma="mi" problem="L_w_su" psource="L_sp" lemma="meu" pos="DP1FSS">mi</tok> <tok id="w-182" lemma="opinión" pos ="NCFS000">opinión</tok><tok id="w-183" lemma="," pos="Fc">,</tok> <tok form="--" id="w-184" lemma="--" pos="Fg">na</tok> <add><tok form="a" id="w-185" lemma="o" pos="DA0F50">a</tok></add> <tok id="w-186" lemma="cociña" pos="NCFS000">cociña</tok> <tok id="w-187" lemma="cada" pos="DI0NN0">cada</tok> <tok id="w-188" lemma="vez" pos="NCFS000">vez</tok> <lb id="e-26" /><tok form="vai" id="w-189" lemma="ir" pos="VMIP3S0">vai</tok> <tok id="w-190" lemma="ser" pos="VSN0000">ser</tok> <tok id="w-191" dcform="máis" problem="O_ac_om" lemma="máis" pos="RG">mais</tok> <tok id="w-192" lemma="popular" pos="AQ0CS">popular</tok> <tok id="w-193" lemma="porque" pos="CS">porque</tok> <tok id="w-194" lemma="ser" pos="VSIP3S0">é</tok> <tok form="--" id="w-195" lemma="--" pos="Fg">unh</tok> <tok id="w-196" lemma="un" pos="DI0MS0">un</tok> <lb id="e-27"/><tok form ="campo" id="w-197" lemma="campo" pos="NCMS000">campo</tok> <tok id="w-198" lemma="moi" pos="RG">moi</tok> <tok id="w-199"

3. Anotación

En el corpus se identifican con códigos y se corrigen todas las formas y secuencias no estándares, mediante un sistema multicapa de seis niveles lingüísticos: ortográfico, morfológico, léxico, gramatical, sémántico y discursivo. Los códigos describen siempre el tipo de desviación con respecto al estándar (por ejemplo, en el ámbito ortográfico, omisión de acento gráfico [O_ac.om] o, en el ámbito léxico, sustitución de una palabra estándar por una no estándar [L_w_su]), pero en algunos casos se añade un segundo código que identifica el origen de la divergencia: por ejemplo, L_sp para las transferencias léxicas del español.

La anotación y corrección de los tokens se realiza a través de un formulario integrado en TEITOK, mientras que las desviaciones que pueden afectar a secuencias (anotaciones multipalabra) se anotan mediante archivos standoff también directamente en la plataforma.

La lematización (asignación de lema y categoría gramatical) se realiza mediante Freeling con posterior revisión manual. Las formas léxicas no estándares (por ejemplo, *platos*) reciben un lema estándar (*prato*) y un lema original (*plato*).

Token value (w-191): mais		
pform	Transcription (Inner XML)	mais
form	Student final version	
ocform	Orthographic standard	máis
mcform	Morphological standard	
lcform	Lexical standard	
gcform	Grammatical standard	
scform	Semantic standard	
dcform	Discursive standard	
lemma	Standard lemma	máis
olemma	Original lemma	
pos	POS tag (standard)	RG gramaticales
opos	POS tag (original)	
problem	Type of problem	O_ac_om
psource	Source of the problem	
dcorrection	Derived correction	
arg	Connector	

Anotación multipalabra

Error Annotation

Edit an-4

File	Token list	Word value	non porque a unha persona non lle guste a elaboración de ese prato a ti non che guste tamén
type	Type	Linguistic area	Grammar
code	Code	Code	G_str_su
reg	Correction	Corrected form	o feito de que a unha per
Save		cancel	
delete segment			

lemma	Standard lemma	prato
olemma	Original lemma	plato
pos	POS tag (standard)	NCMS000 gramaticales
opos	POS tag (original)	gramaticales
problem	Type of problem	L_w_su
psource	Source of the problem	L_sp
dcorrection	Derived correction	
arg	Connector	

Además, a cada texto se le asignan, también mediante un formulario, diferentes metadatos, entre los que cabe destacar los cuantitativos: número de lemas, palabras, enunciados y párrafos, densidad léxica, media de enunciados por párrafo y de palabras por enunciado y número de palabras del enunciado más corto y del más largo.

4. Visualización

Los textos pueden verse con las formas eliminadas (en gris y tachadas), como en la imagen, o en la versión final, sin las formas suprimidas. Las añadidas se destacan en rojo y las de lectura dudosa con fondo verde.

Opciones de visualización

Texto: **Transcripción completa** **Versión final estudiante** **Estándar ortográfico** **Estándar morfológico** **Estándar léxico**

Estándar gramatical **Estándar semántico** **Estándar discursivo**

Mostrar: **Colores** **Alineación** **<pb>** **<lb>**

Etiquetas: **Lema estándar** **Lema original** **Clase de palabra (estándar)** **Clase de palabra (original)** **Tipo de problema**

Origen del problema **Corrección derivada** **Conector**

A gastronomía nos últimos anos supuxo unha gran popularidade social.

Os cociñeiro famosos da actualidade non teñe máis de cincuenta anos áinda que sempre hai excepcións. O auxe que tivo o mundo da gastronomía vén impulsado polo achegamento a tecnoloxía e sobre todo ao internet. Grazas ao internet podemos buscar todo tipo de recetas e ata os cociñeiro suben vídeos explicando como se fai comidas determinadas, entón desta maneira impulsa a xente a interesarse por este mundo. Outras das cousas polo que se ve influída influída a comida é cando alguén deixa a súa casa e ten que valerse por si só, cociñando para el ou ela mesmo e inventando recetas ata conseguir unha variedade de comidas

O **cad** Aínda que os programas de televisión, os novos como "Master chef" poden ser unha moda pasaxeira porque ano tras ano céntranse máis nos problemas entre os compañeiros do programa pero tamén supuxo a popularidade da cociña. Pero hai programas que nunca van a pasar de moda como "Arguiñano"

En mi opinión, na **a** cociña cada vez vai ser mais popular porque é **unha** un campo moi amplio a novidade e vánse descubrir cosas novas e a xente vaille gustar cada vez máis a gastronomía

Los textos también pueden visualizarse con las correcciones de cada nivel destacadas en distintos colores. Cada capa hereda las del nivel anterior, de modo que en el último, el discursivo, se visualizan todas ellas, como en la imagen. Al poner el cursor sobre un token se ven las anotaciones asignadas.

Opciones de visualización

Texto: **Transcripción completa** **Versión final estudiante** **Estándar ortográfico** **Estándar morfológico** **Estándar léxico**

Estándar gramatical **Estándar semántico** Estándar discursivo

Mostrar: **Colores** Alineación <pb> <lb>

Etiquetas: Lema estándar Lema original Clase de palabra (estándar) Clase de palabra (original) Tipo de problema

Origen del problema Corrección derivada Conektor

A gastronomía nos últimos anos **tivo** unha gran popularidade social.

Os cociñeiro famosos da actualidade non **teñen** máis de cincuenta anos, aínda que sempre hai excepcións que tivo o mundo da gastronomía vén impulsado polo achegamento á tecnoloxía e sobre todo **a Internet**. **Internet** podemos buscar todo tipo de **receitas** e ata os cociñeiro **soben** vídeos explicando como se fai coas determinadas; entón, desta maneira, impulsa a xente a interesarse por este mundo. **Outra das circunstancias** que se ve influída a comida **encontrámola** cando alguén deixa a súa casa e ten que valerse por si só, cociñársela mesmo e inventando **receitas** ata conseguir unha **certa** variedade de comidas.

Aínda que os programas de televisión, os novos como "**MasterChef**", poden ser unha moda pasaxeira porque no ano **se centran** máis nos problemas entre os compañeiros do programa, tamén **supuxeron** a popularidade. Pero hai programas que nunca van pasar de moda como "**Arguiñano**".

Na miña opinión, **a** cociña cada vez vai ser **máis** popular porque é un campo moi **aberto** á novedade e **vanse** descubrir **cousas** novas e **á** xente vaille gustar cada vez más a gastronomía.

supuxo

Estándar semántico	tivo
Lema estándar	supor
Clase de palabra (estándar)	Verbo (VMIS350) Main; indicative; past; third; singular
Tipo de problema	S_w_su

5. Consultas

El corpus ofrece un buscador que, entre otras posibilidades, permite realizar b squedas por los c digos que identifican las formas desviantes del est ndar. Tambi n se pueden usar los metadatos como filtros.

Búsqueda en el corpus

Búsqueda: Buscar constructor de consultas | visualizar | opciones CQL:

Constructor de Consultas

Búsqueda de texto

Versión final estudiante	igual a <input type="button" value="▼"/>	<input type="text"/>
Estándar ortográfico	igual a <input type="button" value="▼"/>	<input type="text"/>
Estándar morfológico	igual a <input type="button" value="▼"/>	<input type="text"/>
Estándar léxico	igual a <input type="button" value="▼"/>	<input type="text"/>
Estándar gramatical	igual a <input type="button" value="▼"/>	<input type="text"/>
Estándar semántico	igual a <input type="button" value="▼"/>	<input type="text"/>
Estándar discursivo	igual a <input type="button" value="▼"/>	<input type="text"/>
Tipo de problema	[seleccionar] <input type="button" value="▼"/>	
Origen del problema	[seleccionar] <input type="button" value="▼"/>	
Clase de palabra (estándar)	[seleccionar] <input type="button" value="▼"/>	
Clase de palabra (original)	[seleccionar] <input type="button" value="▼"/>	
Lema estándar	igual a <input type="button" value="▼"/>	<input type="text"/>
Lema original	igual a <input type="button" value="▼"/>	<input type="text"/>
Conector	[seleccionar] <input type="button" value="▼"/>	

Añadir token

Búsqueda de documentos

Título	<input type="text"/>
Tema	[seleccionar] <input type="button" value="▼"/>
Convocatoria del examen	[seleccionar] <input type="button" value="▼"/>
Comisión delegada	[seleccionar] <input type="button" value="▼"/>
Número de palabras	<input type="text"/>
Número de lemas	<input type="text"/>
Densidad léxica	<input type="text"/>
Número de enunciados	<input type="text"/>
Palabras por enunciado	<input type="text"/>
Palabras enunciado más largo	<input type="text"/>
Palabras enunciado más corto	<input type="text"/>
Número de párrafos	<input type="text"/>
Enunciados por párrafo	<input type="text"/>

Anotaciones multipalabra

Código de error [seleccionar]

Las b usquedas de texto ofrecen una concordancia KWIC con diferentes ordenaciones posibles. Adem s, tras cada concordancia puede realizarse una consulta por frecuencia, que permite obtener datos de distribuci n del elemento buscado de acuerdo con diferentes criterios.

Búsqueda en el corpus

Búsqueda CQL: [olemma = "abuelo"] within text

Buscar constructor de consultas | visualizar | opciones

3 resultados • ipm: 10.92

Texto: [Transcripción completa](#) [Versión final estudiante](#) [Estándar morfológico](#) [Estándar léxico](#) [Estándar discursivo](#)

Etiquetas: [Lema estándar](#) [Lema original](#) [Clase de palabra \(estándar\)](#) [Tipo de problema](#) [Origen del problema](#) [Corrección derivada](#)

Contexto os típicos pratos da **abuela** , dende o | meu

Contexto selecta, como di meu **abuelo** feita con pinzas,

Contexto nenos, que | son os **abuelos** e eso por unha

Descargar resultados • guardar esta consulta

Consultas sobre frecuencia

Frecuencia por: [seleccionar] ▾