

CORTEGAL. Corpus de textos galegos escritos por estudantes no ámbito académico

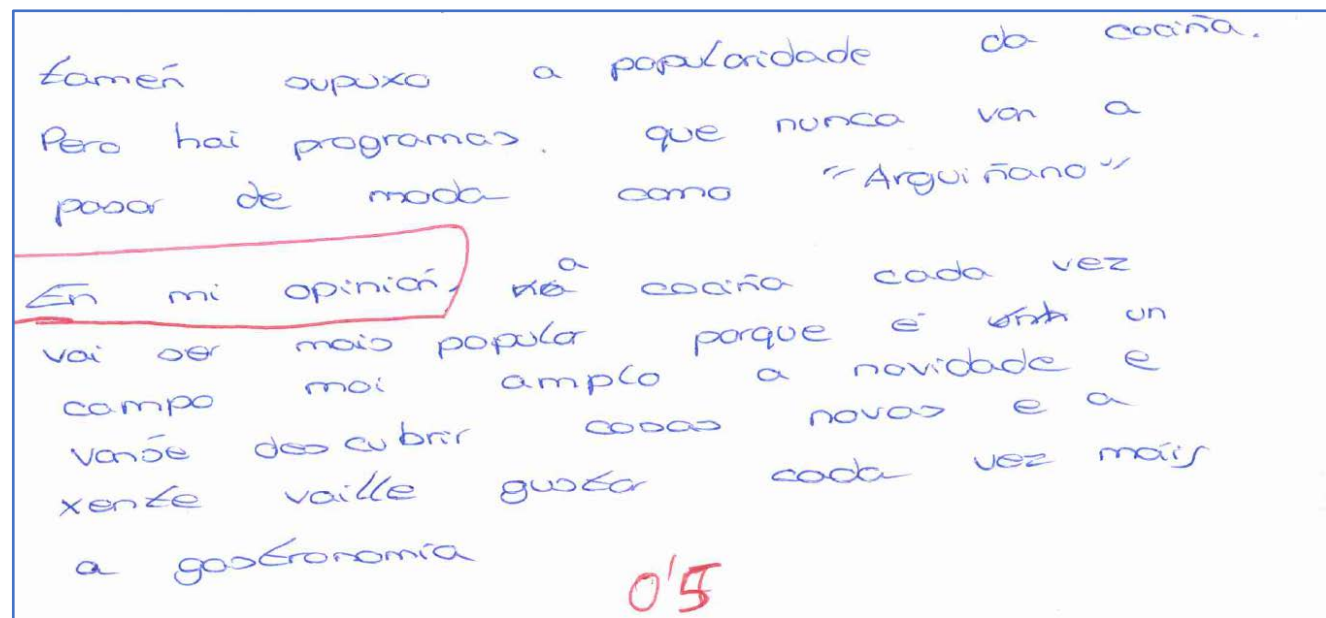
María Álvarez de la Granja
Instituto da Lingua-USC
maria.alvarez.delagranja@usc.gal



I Xeira CLARIAH-GAL
Santiago de Compostela
2 de maio de 2024

1. Presentación

- ✓ 1000 textos argumentativos dos exames de lingua galega das probas ABAU (curso 2016-2017), con anotación das formas non estándares.
- ✓ Obxectivos:
 - Coñecer as características da produción escrita en lingua galega do alumnado de Galicia ao finalizar a educación secundaria.
 - Servir como ferramenta de axuda na aula.



2. Transcrición

- ✓ Transcrición e anotación en TEITOK.
- ✓ Transcrición e etiquetado das formas riscadas polo/a estudante, das claramente engadidas a e das de lectura dubidosa.

```
>Arguiñano</tok><tok id="w-179" lemma="mais" pos="RG">mais</tok><tok id="w-219" dcform="." problem="D_ac_om">.</tok><p><tok id="p-4"><tok id="w-180" gcform="Na" problem="G_det_om" psouce="G_sp" lemma="en" pos="SP">En</tok><tok id="w-181" lcform="miña" olemma="mi" problem="L_w_su" psouce="L_sp" lemma="meu" pos="DP1F5S">mi</tok><tok id="w-182" lemma="opinión" pos="NCFS000">opinión</tok><tok id="w-183" lemma="a" pos="F">a</tok><tok id="w-184" lemma="e" pos="Fg">e</tok><del><tok form="a" id="w-185" lemma="o" pos="DA0F50">o</del><tok id="w-186" lemma="cocina" pos="NCFS000">cocina</tok><tok id="w-187" lemma="cada" pos="D10N00">cada</tok><tok id="w-188" lemma="vez" pos="NCFS000">vez</tok><tok id="w-189" lemma="e" pos="F">e</tok><tok id="w-190" lemma="ser" pos="VSN0000">ser</tok><tok id="w-191" ocform="mais" problem="O_ac_om" lemma="mais" 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="VSP350">é</tok><del><tok form="a" id="w-195" lemma="a" pos="Fg">a</del><tok id="w-196" lemma="un" pos="D10NS0">un</tok><tok id="w-197" lemma="campo" id="w-197" lemma="campo" pos="NCMS000">campo</tok><tok id="w-198" lemma="moi" pos="RG">moi</tok><tok id="w-199" lemma="amplo" id="w-199" lemma="amplo" pos="NCMS000">amplo</tok><tok id="w-200" lemma="a" pos="F">a</tok><tok id="w-201" lemma="xente" pos="NCMS000">xente</tok><tok id="w-202" lemma="vaile" pos="VSP350">vaile</tok><tok id="w-203" lemma="gustar" pos="VSP350">gustar</tok><tok id="w-204" lemma="cada" pos="D10N00">cada</tok><tok id="w-205" lemma="vez" pos="NCFS000">vez</tok><tok id="w-206" lemma="mais" pos="RG">mais</tok></p></pre>
```

3. Anotación

- ✓ Sistema multicapa de seis niveis lingüísticos: ortográfico, morfolóxico (flexión), léxico, gramatical (sintáctico), semántico e discursivo.
- ✓ Códigos que describen o tipo de desviación respecto ao estándar: omisión de acento gráfico (O_ac_om); substitución dunha palabra estándar por unha non estándar (L_w_su)...
- ✓ Nalgúns casos, códigos que identifican a orixe da diverxencia: transferencias léxicas do español (L_sp); flexión dialectal (M_gal)...
- ✓ Anotación e corrección dos tokens a través dun formulario integrado en TEITOK e anotacións multipalabra mediante arquivos *standoff*.
- ✓ Lematización (lema e categoría gramatical) mediante *Freeling* con posterior revisión manual.
- ✓ Asignación ás formas léxicas non estándares (p.e. *platos*) dun lema estándar (*prato*) e dun lema orixinal (*plato*).

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

Anotación multipalabra	
Error Annotation	
Edit an-4	
type	Type
code	Code
mg	Correction
mg	Corrected form
token	Token
lemma	Standard lemma
olemma	Original lemma
pos	POS tag (standard)
opos	POS tag (original)
problem	Type of problem
psource	Source of the problem
dcorrection	Derived correction
arg	Connector

- ✓ Asignación de metadatos cuantitativos: número de lemas, palabras, enunciados e parágrafos, densidade léxica, media de enunciados por parágrafo e de palabras por enunciado e número de palabras do enunciado máis curto e do máis longo.

4. Visualización

- ✓ Distintas posibilidades de visualización:
 - Coas formas eliminadas (en gris e riscadas), como na imaxe, ou na versión final, sen as formas suprimidas. As engadidas destacan en vermello e as de lectura dubidosa con fondo verdoso.

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 cocinheiros famosos da actualidade non teñen máis de cincuenta anos aínda 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 cocinheiros 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 a comida é cando alguén deixa a súa casa e ten que valerse por si só, cocinando para el mesmo e inventando recetas ata conseguir unha variedade de comidas.

En mi opinión, na a cocinha cada vez vai ser máis popular porque é un campo moi amplo a novidade e vánse descubrir cousas novas e a xente vaile gustar cada vez máis a gastronomía.

- Coas estandarizacións de cada nivel destacadas en distintas cores: no último nivel, o discursivo, visualízanse todas elas, como na imaxe.

Opciones de visualización

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

Estándar gramatical | Estándar semántico | Estándar discursivo

Mostrar: Cores | Alineación | <pb> | <lb>

Anotación: Lema estándar | Lema original | Clase de palabra (estándar) | Clase de palabra (original)

Tipo de desviación do estándar | Fonte de formas non estándar | Corrección derivada | Conector

A gastronomía nos últimos anos tivo unha gran popularidade social.

Os cocinheiros famosos da actualidade non teñen máis de cincuenta anos, aínda que sempre hai excepcións. O auxe que tivo o mundo da gastronomía vén impulsado polo achegamento á tecnoloxía e sobre todo a Internet. Grazas a Internet podemos buscar todo tipo de receitas e ata os cocinheiros soben vídeos explicando como se fai comidas determinadas; entón, desta maneira, impulsa a xente a interesarse por este mundo. Outra das circunstancias polas que se ve influída a comida encontrámola cando alguén deixa a súa casa e ten que valerse por si só, cocinando para el 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 ano tras ano se centran máis nos problemas entre os compañeiros do programa, tamén supuxeron a popularidade da cocinha. Pero hai programas que nunca van pasar de moda como "Arguiñano".

Na miña opinión, a cocinha cada vez vai ser máis popular porque é un campo moi aberto á novidade e vanse descubrir cousas novas e a xente vaile gustar cada vez máis a gastronomía.

5. Consultas

- ✓ Entre outras, buscas polos códigos das formas desviantes do estándar.
- ✓ Uso dos metadatos como filtros de documentos.

Buscas no corpus

Consulta CQL:

Xerador de consultas

Búsqueda de texto	Búsqueda de documentos
Versión final	Título
Estándar ortográfico	Clase de palabras
Estándar morfológico	Clase de palabras (original)
Estándar léxico	Clase de palabras (estándar)
Estándar gramatical	Clase de palabras (original)
Estándar semántico	Clase de palabras (estándar)
Estándar discursivo	Clase de palabras (original)
Tipo de desviación do estándar	Clase de palabras (estándar)
Fonte de formas non estándar	Clase de palabras (original)
Clase de palabra (estándar)	Clase de palabras (original)
Clase de palabra (original)	Clase de palabras (estándar)
Lema estándar	Clase de palabras (original)
Lema original	Clase de palabras (estándar)
Conector	Clase de palabras (original)

Engade palabra

Buscar Cancelar Ajudar

- ✓ Concordancia KWIC con diferentes ordenacións posibles.
- ✓ Consulta por frecuencia, que permite obter datos de distribución do elemento buscado de acordo con diferentes criterios.

Distribución no corpus

Consulta Tipo de desviación do estándar = (*,?)L_ac_su(.,*)?

Criterio de agrupamento	Lema orixinal
Total	24
Tamaño de referencia	274617

Gráfica: Táboa

élite	7	25.49	29.17
textil	5	18.21	20.83
canón	4	14.57	16.67
período	3	10.92	12.5
colón	2	7.28	8.33
Nóbel	1	3.64	4.17
nóbel	1	3.64	4.17
vértice	1	3.64	4.17