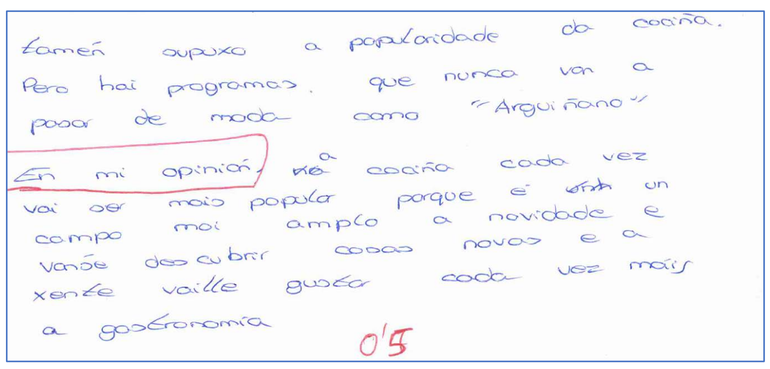
CORTEGAL. Corpus de textos galegos escritos por estudantes no ámbito académico

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

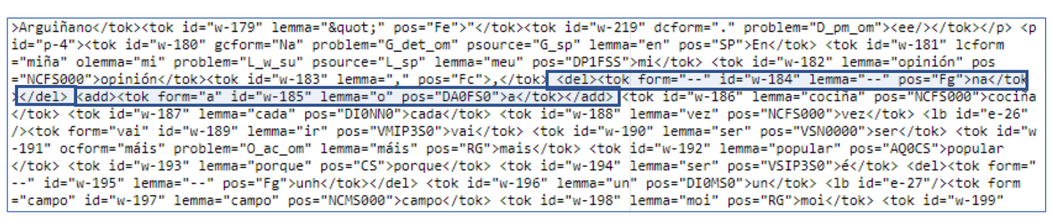
1. Presentación

- ✓ 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.



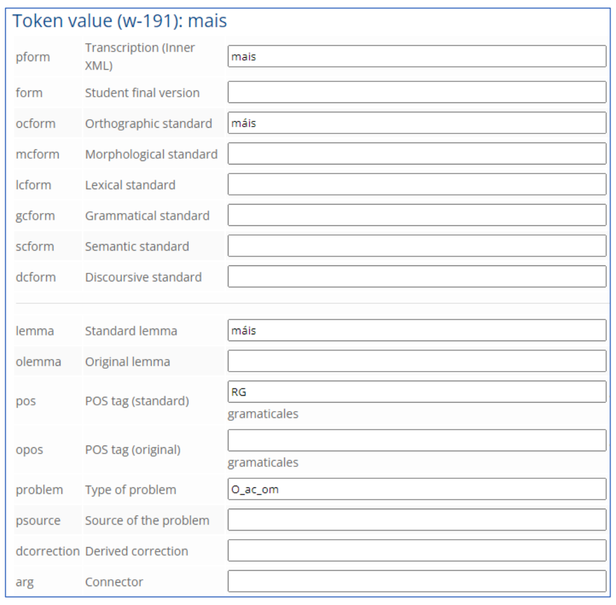
2. Transcrición

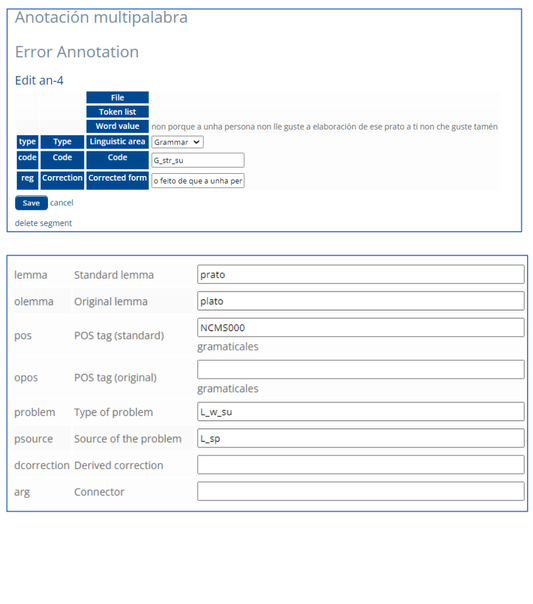
- ✓ Transcrición e anotación en TEITOK.
- ✓ Transcrición e etiquetado das formas riscadas polo/a estudante, das claramente engadidas a e das de lectura dubidosa.



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*).

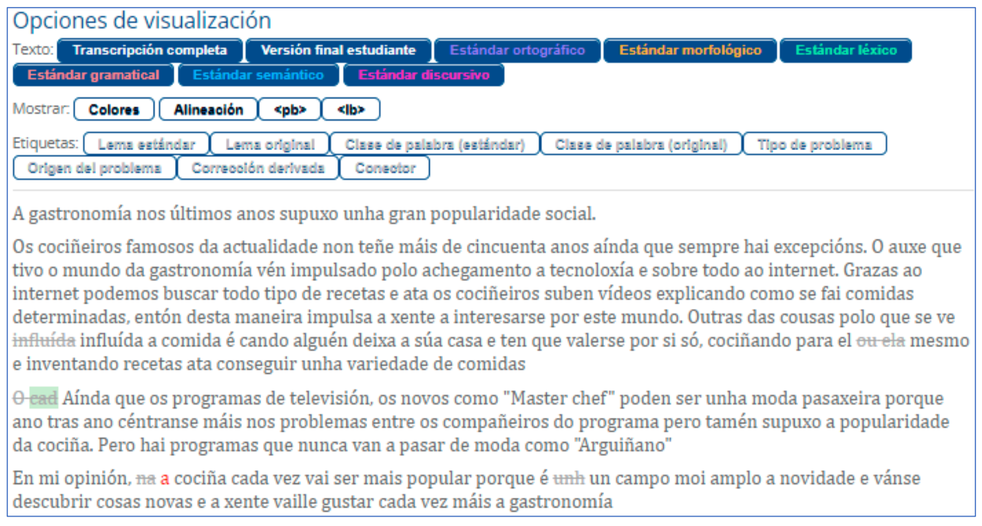




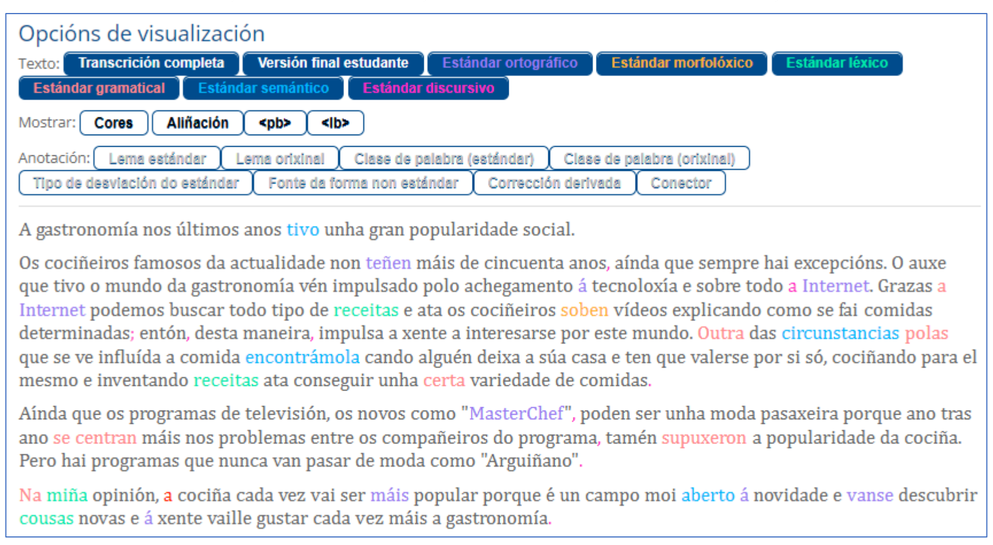
- ✓ Asignación de metadatos cuantitativos: número de lemas, palabras, enunciados e parágrafos, densidade léxica, media de enunciados por parágrafo e de palabras por enunciado e número de palabras do enunciado máis curto e do máis longo.

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.

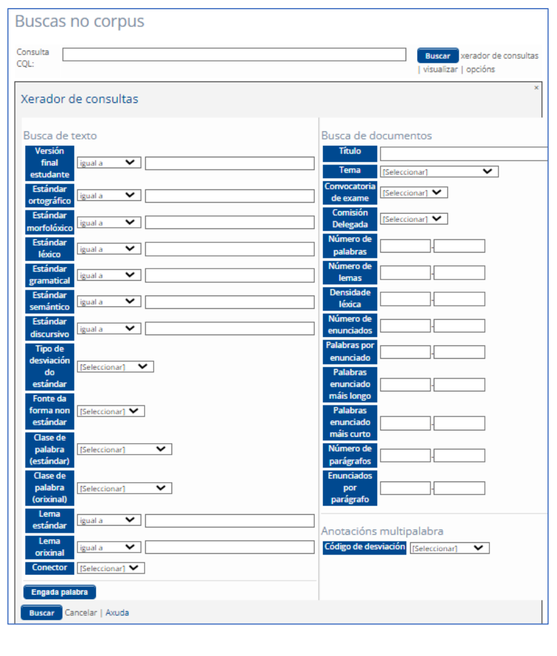


- Coas estandarizacións de cada nivel destacadas en distintas cores: no último nivel, o discursivo, visualízanse todas elas, como na imaxe.

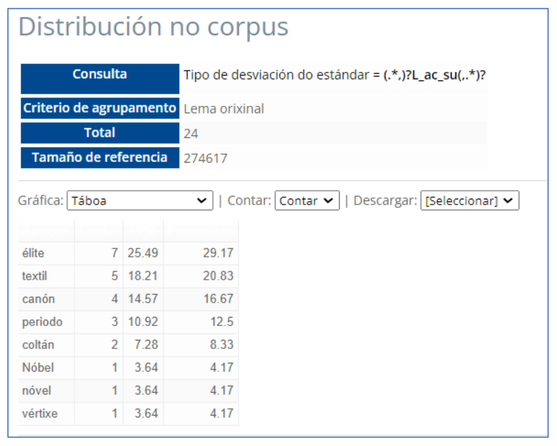


5. Consultas

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



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



Actividade realizada en colaboración con
GOBIERNO DE ESPAÑA MINISTERIO DE CULTURA
MINISTERIO DE CIENCIA E INNOVACIÓN UNIÓN EUROPEA FONDO EUROPEO DE DESARROLLO REGIONAL AGENCIA ESTATAL DE INVESTIGACIÓN
XUNTA DE GALICIA CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
INSTITUTO DA LINGUA GALEGA (ILG)