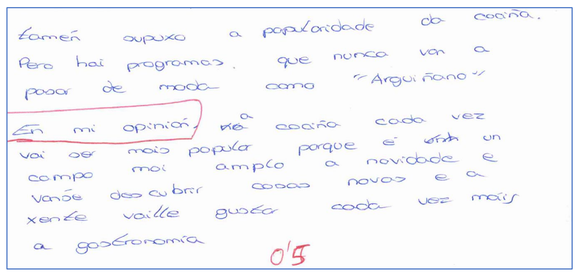
CORTEGAL. Corpus de textos gallegos escritos por estudiantes en el ámbito académico

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

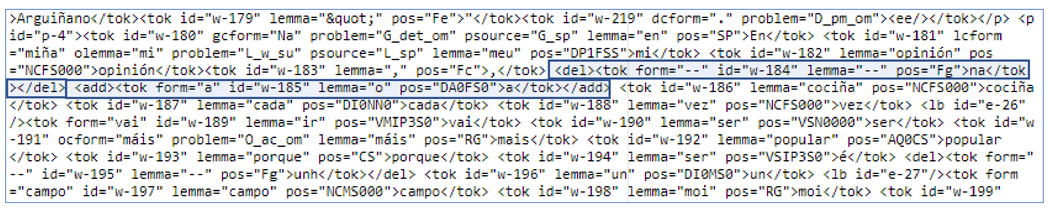
1. Presentación

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.



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.

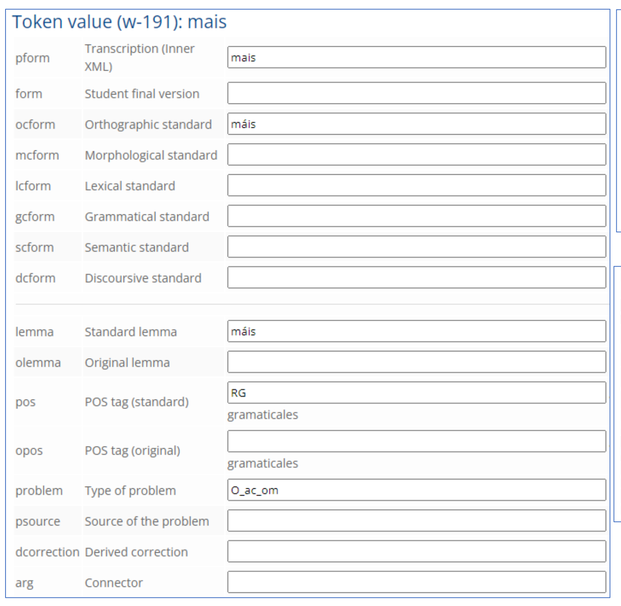


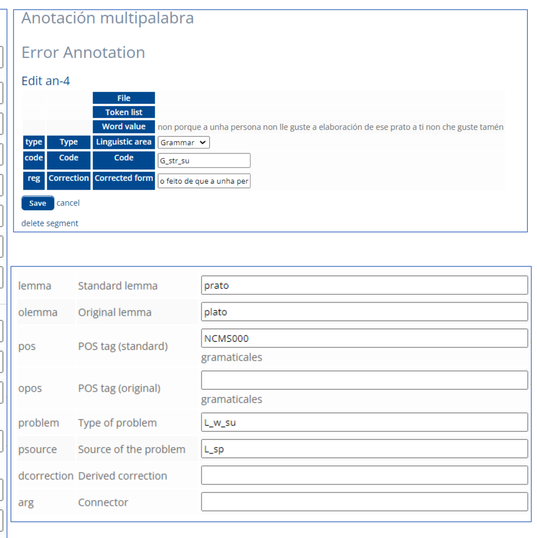
3. Anotación

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*).

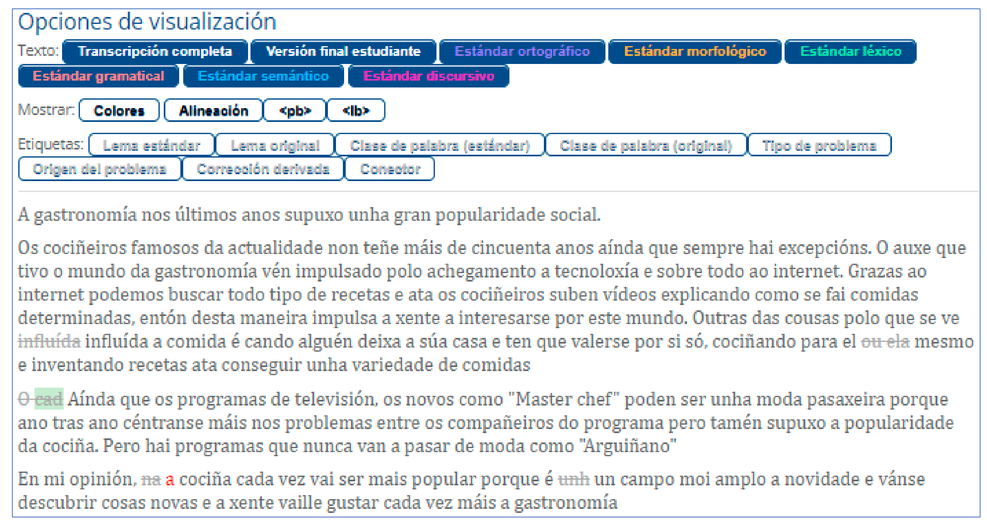




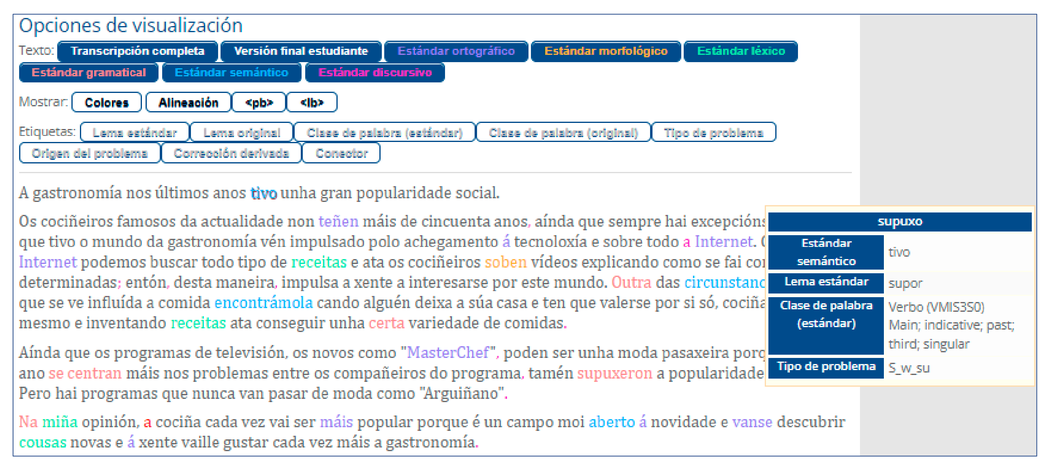
Además, a cada texto se le asignan, también mediante un formulario, diferentes metadatos, entre los que cabe destacar los cuantitativos: número de lemas, palabras, enunciados y párrafos, densidad léxica, media de enunciados por párrafo y de palabras por enunciado y número de palabras del enunciado más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

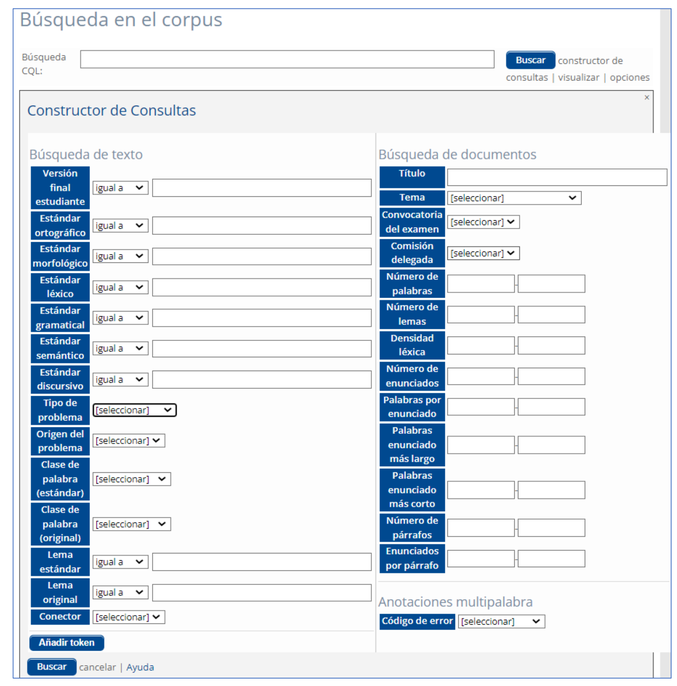


Los textos también pueden visualizarse con las correcciones de cada nivel destacadas en distintos colores. Cada capa hereda las del nivel anterior, de modo que en el último, el discursivo, se visualizan todas ellas, como en la imagen. Al poner el cursor sobre un token se ven las anotaciones asignadas.

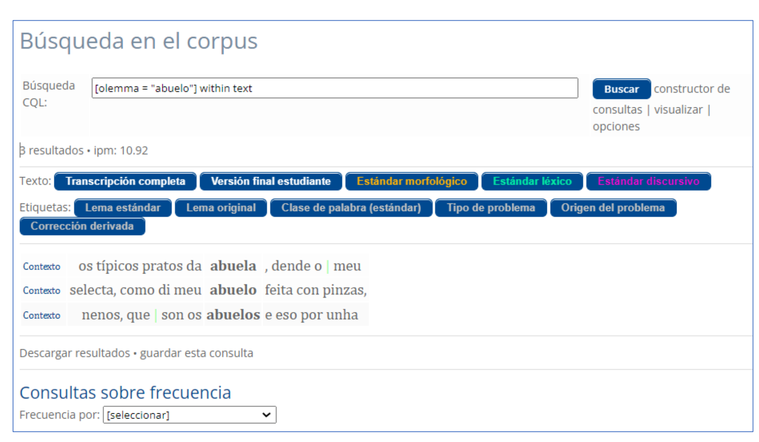


5. Consultas

El corpus ofrece un buscador que, entre otras posibilidades, permite realizar búsquedas por los códigos que identifican las formas desviantes del estándar. También se pueden usar los metadatos como filtros.



Las búsquedas de texto ofrecen una concordancia KWIC con diferentes ordenaciones posibles. Además, tras cada concordancia puede realizarse una consulta por frecuencia, que permite obtener datos de distribución del elemento buscado de acuerdo con diferentes criterios.



Proyecto PGC2018-096069-B-I00, financiado por MCIN/AEI/10.13039/501100011033/ y por FEDER "Una manera de hacer Europa"