

COTWEETGAL: UN NOVO RECURSO PARA O ESTUDO DO GALEGO NAS REDES SOCIAIS

NOELIA ESTÉVEZ RIONEGRO

noelia.rionegro@usc.gal

Instituto da Lingua Galega, Universidade de Santiago de Compostela

A consolidación das redes sociais como espazos de interacción comunicativa transformou as formas de uso da lingua e abriu novas posibilidades para a investigación lingüística e sociolingüística. Neste contexto, os corpus de comunicación dixital constitúen ferramentas fundamentais para analizar fenómenos relacionados coa variación lingüística, o multilingüismo, a innovación léxica, as prácticas discursivas e a construción de identidades en contornos en liña. Porén, no caso galego seguen sendo escasos os recursos especificamente deseñados para o estudo da comunicación nas redes sociais desde unha perspectiva lingüística ampla e multidimensional.

Esta comunicación presenta *COTWEETGAL: Corpus de tweets publicados en Galicia*, un recurso elaborado no Instituto da Lingua Galega que recompila 72.018 chíos xeolocalizados en Galicia, publicados entre os anos 2019 e 2021. O corpus foi construído a partir de datos recuperados mediante a API de Twitter e inclúe información temporal, xeográfica e comunicativa, así como diferentes indicadores de interacción, entre eles o número de *likes*, *retweets* e *hashtags*. Os datos foron sometidos a procesos de limpeza, anonimización e estruturación que facilitan a súa explotación con fins científicos.

Un dos principais valores de COTWEETGAL reside no seu carácter multilingüe. A identificación lingüística realizouse mediante un sistema híbrido que combina varios clasificadores automáticos e unha revisión manual selectiva, o que permitiu diferenciar con maior precisión linguas próximas como o galego, o castelán e o portugués. Os resultados mostran a presenza de 27 linguas, aínda que o corpus está dominado polo castelán e o galego, seguidos do inglés e do portugués, ademais de numerosas combinacións plurilingües e mensaxes constituídas exclusivamente por emoticonas ou outros elementos non verbais.

O recurso incorpora, ademais, unha primeira capa de anotación ortográfica e codificación segundo o estándar TEI-XML, que permite representar elementos característicos da comunicación dixital, como *hashtags*, emoticonas e determinados

usos gráficos expresivos. Esta estrutura constitúe a base para futuros procesos de enriquecemento lingüístico e garante a interoperabilidade do corpus con outras infraestruturas de investigación.

A comunicación describirá o deseño metodolóxico do corpus, as súas características principais e o seu potencial para o estudo do galego nas redes sociais. Así mesmo, mostraranse exemplos de explotación dos datos relacionados coa distribución das linguas, o contacto lingüístico, a escritura dixital e as prácticas comunicativas desenvolvidas na Galicia contemporánea. COTWEETGAL aspira, deste xeito, a converterse nun recurso de referencia para a investigación sobre comunicación dixital, multilingüismo e variación lingüística no ámbito galego.