II XEIRA CLARIAH-GAL
9 de maio de 2025
Facultade de Filoloxía da Universidade de Santiago de Compostela

PROXECTOS E RECURSOS DO INSTITUTO DE INVESTIGACIÓN LINGUA DA UNIVERSIDADE DE VIGO

JAVIER PÉREZ GUERRA

iLingua, Universidade de Vigo

Nesta presentación damos conta de recursos e proxectos informáticos e computacionais desenvolvidos por membros do novo Instituto Universitario de Investigación Lingua (iLingua) da Universidade de Vigo.

(i) Bases de datos:

- ADESSE 'Base de datos de verbos, alternancias de diátesis y esquemas sintáctico-semánticos del español' (coordinador: José María García-Miguel) https://adesse.uvigo.es. ADESSE é unha base de datos de verbos e construcións verbais do español coa análise sintáctico-semántica típica dun corpus, que permite ofrecer para cada verbo unha completa caracterización sintáctico-semántica, coas súas alternancias de diátese xunto coas frecuencias relativas a relacións semánticas similares.
- ECEG 'Eighteenth-Century English Grammars database' (coordinadora: Nuria Yáñez) https://eceg.iatext.ulpgc.es. ECEG é unha base de datos compilada no contexto da estandarización escrita e oral no século XVIII, con metadatos bibliográficos e biográficos sobre 323 gramáticas do inglés e 275 autores/as.
- ECEP 'The Eighteenth-Century English Phonology database' (coordinadora: Nuria Yáñez e equipo do Digital Humanities Institute de Sheffield). Base de metadatos de máis de 21.000 palabras con información fonolóxica (IPA) extraída de 11 dicionarios de pronuncia publicados na segunda metade do século XVIII.
-

(ii) Corpus:

- Lingua de signos española:
  - iSignos 'Interfaz de datos de la Lengua de Signos Española' (coordinadores: Carmen Cabeza e José María García-Miguel) https://isignos.uvigo.es. iSignos é un recurso creado a partir do proxecto RADIS, para a consulta de datos lingüísticos da lingua de signos española. O corpus consiste nun conxunto de 30 gravacións en vídeo de signantes, presentadas xunto ás glosas das mans e maila tradución ao español e ao inglés. O equipo desenvolveu tamén o sistema de anotación gramatical empregado neste corpus (formato ELAN .eaf): 'Annotations for LSE-RADIS corpus' https://zenodo.org/records/10670864
  - UD_Spanish_Sign_Language-LSE 'Universal Dependencies treebank da Lingua de Signos Española' (coordinador: José Mª García-Miguel) https://universaldependencies.org/treebanks/ssp_lse/index.html, https://github.com/UniversalDependencies/UD_Spanish_Sign_Language-LSE. 500 oracións analizadas mediante gramática de dependencias.

- Lingua inglesa e variedades:
  - ICE-GBR, componente xibraltareño do International Corpus of English (consorcio formado por 26 grupos internacionais con sede na Universidade de Zürich) (coordinadora: Elena Seoane). 1 millón de palabras de 20 rexistros diferentes de inglés oral e escrito (desde conversas espontáneas ata debates parlamentarios). Parte escrita dispoñible https://view0.webs.uvigo.es/ice-gibraltar
  - SGibE 'Corpus of Spoken Gibraltar English' (coordinadora: Elena Seoane). Corpus oral de inglés de Xibraltar, espontáneo de falantes xibraltareños do podcast *Street Talk* (en construción).
  - GibPress 'Corpus of Gibraltar Press' (coordinadora: Elena Seoane). Corpus de prensa xibraltareña, con reportaxes e editoriais dos sitios web do *Gibraltar Chronicle* e *Panorama* dos anos 2014 e 2024, permitindo así investigacións diacrónicas. Conta cun sistema automatizado para recuperar e arquivar textos de ambos xornais, garantindo o cumprimento das directrices éticas e legais (en construción).
  - 'The Mary Hamilton Papers' (coordinadora: Nuria Yáñez) https://www.digitalcollections.manchester.ac.uk/collections/maryhamilton/1. Edición dixital de 3.200 elementos e preto de 18.000 imaxes dixitalizadas de ego-documentos (diarios, cartas privadas e manuscritos), dispoñible no repositorio Manchester Digital Collections, preparada como corpus lingüístico na interface CQPweb, con anotación morfolóxica e semántica. É tamén unha edición crítica en tanto que inclúe anotacións manuais de carácter sociocultural, histórico, literario e biográfico.
  - ARCHER 'A Representative Corpus of Historical English Registers' (coordinadora: Nuria Yáñez) https://www.projects.alc.manchester.ac.uk/archer. ARCHER é un corpus histórico de rexistros textuais en lingua inglesa dispoñible na interface CQPweb, con anotación morfolóxica e semántica. Abrangue catro séculos e 12 rexistros de carácter formal e informal, nas variedades de inglés británico e americano, cun total de 3,3 millóns de palabras.
  - 'The APU Writing and Reading Corpus' (coordinadora: Nuria Yáñez e a Universidade de Liverpool) https://datacat.liverpool.ac.uk/2456. APU combina a edición dixital, con imaxes orixinais e corpus lingüísticos, con transcricións en versión diplomática e normalizada, e anotación morfolóxica e semántica. Corpus elaborado no marco da lingüística educativa na época de Margaret Tatcher, con materiais do ano 1979 e 1988 producidos por estudantes de nivel Year 6 e Year 11 en Reino Unido, cun total de 172.000 palabras.
  - VICOLSE 'Vigo Corpus of Learner Spoken English' (coordinadores: Beatriz Tizón, Javier Pérez). Corpus de aprox. 100.000 palabras de inglés oral narrativo e argumentativo, producido por estudantes universitarios de inglés como lingua estranxeira na Uvigo, con información sociolóxico-cultural das persoas participantes. A plataforma (en construción) facilita acceso aos audios e ás transcricións, seguindo as convencións de LINDSEI.

(iii) Plataformas e portais de consulta:

- RILG 'Recursos Integrados da Lingua Galega' (coordinado polo grupo TALG 'Tecnoloxías e Aplicacións da Lingua Galega', Alexandre Fernández) https://ilg.usc.gal/rilg. O portal integra dende 2006 diversos recursos textuais e léxicos de tecnoloxía lingüística da lingua galega xerados en distintos proxectos, posibilitando así a consulta libre e simultánea de todos eles. Os recursos incorporan dicionarios galegos, corpus de lingua galega e corpus paralelos de tradución.
- NEOTECA (coordinado polo grupo TALG 'Tecnoloxías e Aplicacións da Lingua Galega', Alexandre Fernández) https://ilg.usc.gal/neoteca. Neoteca é o banco de datos sobre neoloxía léxica do galego a partir dos datos obtidos polo Observatorio de Neoloxía do grupo TALG, coa colaboración do SLI. De cada neoloxismo documentado na prensa galega dende 1997 fornécese información contextual, datación, fonte, categoría gramatical e procedemento de formación.
- RERCOR v3.1.2 2018-2025 'Recursos sobre enfermidades raras' (coordinadoras: Elena Sánchez e Tamara Varela) http://www.rercor.org. RERCOR é un portal de recursos lingüísticos multilingües do ámbito biomédico sobre un amplo grupo de enfermidades raras, minoritarias ou orfas, baseado en corpus de máis de 8 millóns de palabras, comparables e paralelos en francés, inglés e español: MYOCOR 2.0 (Corpus de enfermidades neuromusculares), EMCOR (Corpus de enfermidades metabólicas), ERCOR (Corpus de guías de práctica clínica sobre ER), ORPHACOR (Corpus de medicamentos orfos), PRODAPCOR (Corpus de produtos de apoio) e TERAPCOR (Corpus de terapia ocupacional).

(iv) Software:

- NER Buddy. Software de análise automática, semiautomática e asistida de subtítulos intralingüísticos en directo con IA (coordinadores: Pablo Romero, Luis Alonso, grupo GALMA). Empregado e testado polo Parlamento Europeo, Netflix e cadeas públicas e empresas internacionais, e testado con modelos comerciais (GPT) como con modelos open-source (Openchat, Llama, Gemma, Mixtral) adestrados polo equipo co obxectivo de afinar o máximo a precisión, isto é, o nivel de coherencia coas avaliacións realizadas por humanos expertos.