https://nos.gal/

@proxectoNos

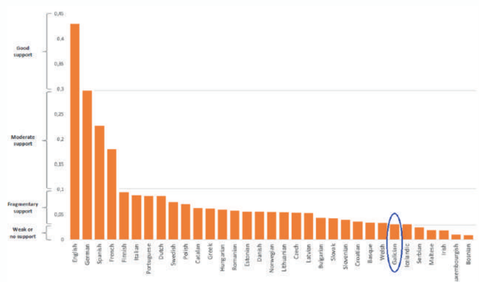
PROXECTO NÓS

Recursos dixitais para a lingua galega

CONTEXTO

O proxecto Nós é unha iniciativa impulsada pola Universidade de Santiago de Compostela e a Xunta de Galicia que busca promover a presenza do galego no ámbito dixital.

Na actualidade, o proxecto forma parte da rede ILENIA, unha iniciativa maior para fornecer tecnoloxías lingüísticas nas linguas oficiais do Estado español.



Estado do soporte tecnolóxico para as linguas europeas.

Fonte: Ramírez Sánchez, J. M., García Mateo, C. (auth.), Giagkou, M., Piperidis, S., Rehm, G., Dunne, J. (eds.). Report on the Galician Language (Deliverable D1.15). ELE, 2022.

OBXECTIVOS INICIAIS

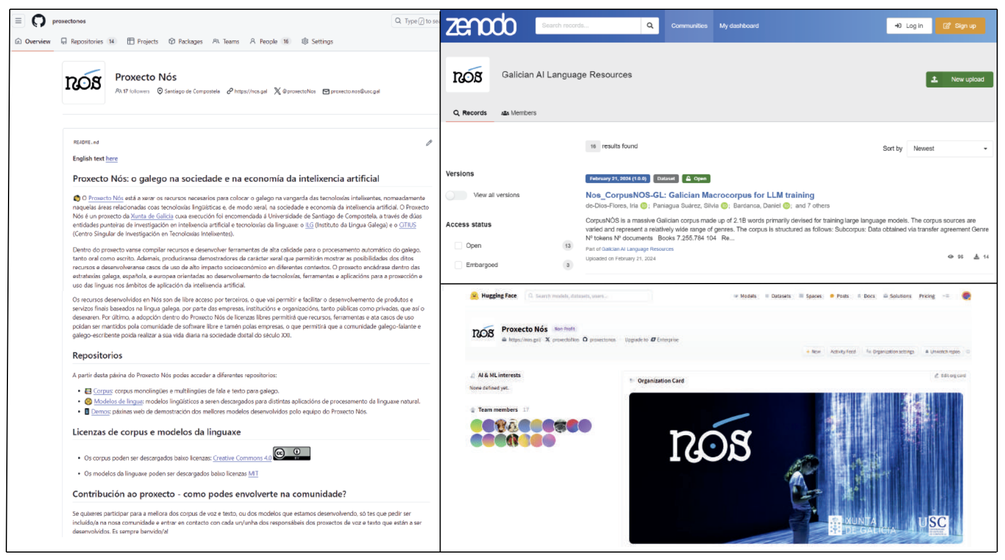
- Crear os medios dixitais necesarios para que o galego prospere como lingua viva na era dixital.
- Desenvolver recursos e ferramentas para o procesamento automático do galego e distribuílos baixo licenzas libres.
- Elaborar demostradores que permitan visibilizar as posibilidades dos recursos.
- Facilitar que empresas e institucións desenvolvan casos de uso.
- Crear un ecosistema galego innovador arredor das tecnoloxías da linguaxe.

RESULTADOS

DATOS					
Corpus	**Tipo**	**Tamaño**	**Utilidade**	**Orixe dos datos**	**Dispoñible en**
Nos_ParlaSpeech_GL	Corpus ASR (aliñado automaticamente)	+1670 h	**Automatic Speech Recognition (ASR)**	Parlamento (2015-2022)	**Zenodo HuggingFace**
Nos_TranscriSpeech_GL	Corpus ASR (aliñado manualmente)	53 h	**ASR**	Conferencias, entrevistas e discursos	**Zenodo HuggingFace**
Nos_GL_CC0	Corpus textual para banco de datos público de voz (Common Voice)	+560.000 frases	**ASR**	Cesión por parte de entidades colaboradoras	**CommonVoice GitHub**
Nos_Celtia_GL	Corpus TTS	25 h	**Text-to-Speech (TTS)**	20.000 frases (extraídas)	**Zenodo**
Nos_CorpusNOS-GL	Macrocorpus	2100 millóns de palabras	**Large Language Models**	Cesión por parte de entidades colaboradoras + corpus web existentes	**Zenodo**
Nos_ES-GL_sin	Corpus paralelo GL-ES	35 millóns de frases	**Machine Translation (MT)**	Parlamento europeo + subtitulado de cinema + OPUS	**Zenodo**
Nos_EN-GL_sin	Corpus paralelo GL-EN	29 millóns de frases	**MT**	Parlamento europeo + subtitulado de cinema + OPUS	**Zenodo**
Nos_MeteoGalicia-GL	Corpus Data-To-Text	3000 rexistros	**Data-To-Text**	Meteogalicia	**Zenodo**
Nos_Conversacional_GL	Corpus conversacional	+553.000 frases	**Chatbots**	CORGA (audiovisual e narrativa) e subtitulado	**Uso exclusivo para investigación**

MODELOS			
Modelo	**Utilidade**	**Arquitectura**	**Dispoñible en**
proxectonos/Nos_TTS-celtia-vits-graphemes	**TTS**	VITS	**HuggingFace**
proxectonos/Nos_TTS-sabela-vits-phonemes	**TTS**	VITS	**HuggingFace**
proxectonos/Nos_ASR-wav2vec2-large-xlsr-53-gl-with-lm	**ASR**	wav2vec2	**HuggingFace**
proxectonos/Nos_D2T-gl	**Data-To-Text Generation**	Seq-to-seq	**HuggingFace**
proxectonos/Cerebras-1.3B-GL	**Text Generation**	Seq-to-seq	**HuggingFace**
proxectonos/FLOR-1.3B-GL	**Text Generation**	Seq-to-seq	**HuggingFace**
proxectonos/Nos_MT-OpenNMT-en-gl	**MT**	Seq-to-seq	**HuggingFace**
proxectonos/Nos_MT-OpenNMT-ca-gl	**MT**	Seq-to-seq	**HuggingFace**
proxectonos/Nos_MT-OpenNMT-eu-gl	**MT**	Seq-to-seq	**HuggingFace**
proxectonos/Nos_MT-OpenNMT-es-gl	**MT**	Seq-to-seq	**HuggingFace**
proxectonos/Nos_MT-OpenNMT-gl-en	**MT**	Seq-to-seq	**HuggingFace**
proxectonos/Nos_MT-OpenNMT-gl-es	**MT**	Seq-to-seq	**HuggingFace**

RECURSOS EN REPOSITORIOS ABERTOS



DEMOSTRADORES			
Tipo	**Utilidade**	**Corpus**	**Arquitectura**
Interface	**ASR**	OpenSLR77	wav2vec2
Interface	**TTS**	Nos_Celtia-GL	VITS
Interface	**MT**	Corpus paralelos	Seq-to-seq
Prototipo	**Data-To-Text**	Corpus data-to-text	Seq-to-seq

Tradutor **TTS** **ASR**

TRABALLO FUTURO

- Aumentar a cantidade e calidade dos corpus actuais.
- Elaborar modelos monolingües e multilingües de voz e texto.
- Xerar datos anotados de calidade.
- Identificar casos de uso de alto impacto e fomentar a transferencia tecnolóxica.

REFERENCIAS

Iria de-Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, José Ramom Pichel, Marcos García, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín-Diz, Manuel González González, Senén Barro, and Xosé Luis Regueira. 2022. The Nós Project: Opening routes for the Galician language in the field of language technologies. In Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference, pages 52–61, Marseille, France. European Language Resources Association.

Iria de-Dios-Flores, Silvia Paniagua Suárez, Cristina Carbajal Pérez, Daniel Bardanca Outeiriño, Marcos Garcia and Pablo Gamallo. 2024. CorpusNÓS: A massive Galician corpus for training large language models. Proceedings of the 16th International Conference on Computational Processing of Portuguese - ACL Anthology (Volume 1), 593-599

Carmen Magariños, Alp Öktem, Antonio Moscoso Sánchez, Marta Vázquez Abuín, Noelia García Díaz, Adina Ioana Vladu, Elisa Fernández Rei, and María Baqueiro Vidal. Nós-TTS: a Web User Interface for Galician Text-to-Speech. Best Demo Award at PROPOR 2024.

Adina Ioana Vladu, Iria de Dios-Flores, Carmen Magariños, John E. Ortega, Jose Ramom Pichel, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín, Manuel González González, Senen Barro, and Xosé Luís Regueira. 2022. Proxecto Nós: Artificial intelligence at the service of the Galician language. In SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, A Coruña, Spain.

Financiado por la Unión Europea NextGenerationEU

Este proyecto con referencia 2022/TL22/00215336 está financiado por el Ministerio de Transformación Digital y por el Plan de Recuperación, Transformación y Resiliencia - Financiado por la Unión Europea – NextGenerationEU