

Proxecto Nós: o galego na vangarda das tecnoloxías intelixentes

XORNADA USUARIOS INFRAESTRUTURA CESGA 2023
HPC, BIG DATA, AI e QUANTUM



NÓS



- **CONTEXTO E OBXECTIVOS**
- **TRABALLO INICIAL**
- **PRÓXIMOS PASOS**
- **TRANSFERENCIA / PARTICIPACIÓN**



CONTEXTO E OBXECTIVOS

Historia



2021

- A Xunta de Galicia, Amtega e a USC asinan o **primeiro convenio** de colaboración para a definición e elaboración do **plan de traballo (2022-2025) do proxecto Nós.**

2022

- A Xunta de Galicia, Amtega e a USC asinan un **segundo convenio** de colaboración para continuar co **desenvolvemento do proxecto Nós.**
- **Presentación pública** do proxecto Nós.

2023

- O proxecto Nós, a través da USC recibe **2M€ do PERTE** Nova Economía da Lingua para o período **2023-2025**, co obxectivo de:
 - Aumentar a cantidade e calidade dos corpus actuais.
 - Elaborar **modelos** de voz e texto, tanto monolingües como **multilingües.**
 - Xerar datos anotados de calidade que permitan adestrar e avaliar os modelos.
 - Identificar casos de uso de alto impacto coa participación da industria, fomentando a transferencia tecnolóxica.
- Colaboración cos outros proxectos beneficiarios do PERTE: **AINA** (Cataluña), **GAITU** (Pais Vasco) e **Vives** (Valencia).

Obxectivos

- 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**.

Liñas de traballo

- ❑ Síntese de voz
- ❑ Recoñecemento da fala
- ❑ Sistemas de diálogo
- ❑ Corrección e avaliación lingüística automáticas
- ❑ Tradución automática
- ❑ Xeración automática de texto
- ❑ Extracción de información
- ❑ Análise de sentimentos e verificación de información



TRABALLO INICIAL CONVENIO XUNTA / USC 2022

ÁREAS DE TRABALLO DURANTE 2022

- **Síntese de voz (TTS)**
- **Recoñecemento da fala (ASR)**
- **Tradución automática**
- **Sistemas de diálogo**
- **Xeración automática de texto**
- Corrección e avaliación lingüística automáticas
- *Extracción de información*
- *Análise de sentimentos e verificación de información*

Resultados 2022

DATOS

Corpus ASR (aliñado)	1750 h	Parlamento (2015-2022) + entrevistas e discursos	Lingua estándar (formal)
Corpus para ASR (datos brutos)	3700 h	TX 2019-2022 + AGO	Lingua estándar + popular
Corpus TTS	30 h	20.000 frases (extraídas)	Lectura
Corpus paralelo galego-español	30 millóns de frases	Parlamento europeo + subtulado de cinema	Lingua estándar formal (parlamento) + informal
Corpus paralelo galego-inglés	25 millóns de frases	Parlamento europeo + subtulado de cinema	Lingua estándar formal (parlamento) + informal
Corpus conversacional	+585.000 frases	Subtitulado	Lingua informal

DEMOSTRADORES

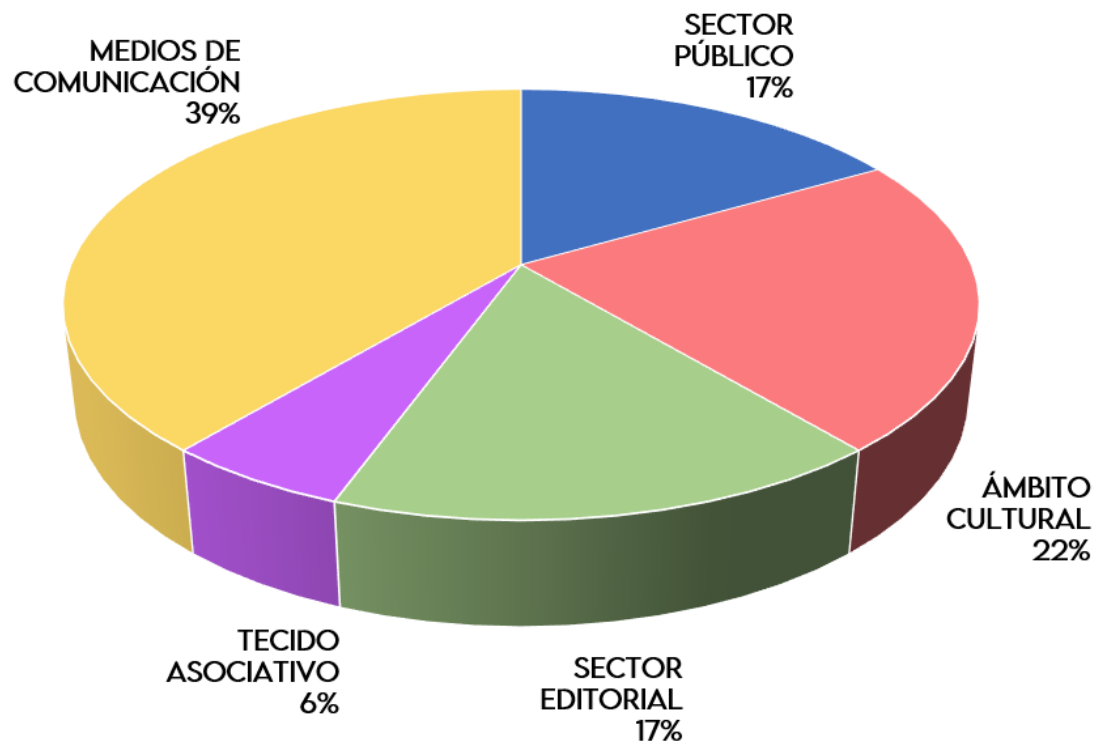
3 Interfaces: **ASR, TTS e Tradución**

1 Prototipo: **Data-to-Text**

ENTIDADES CONTACTADAS

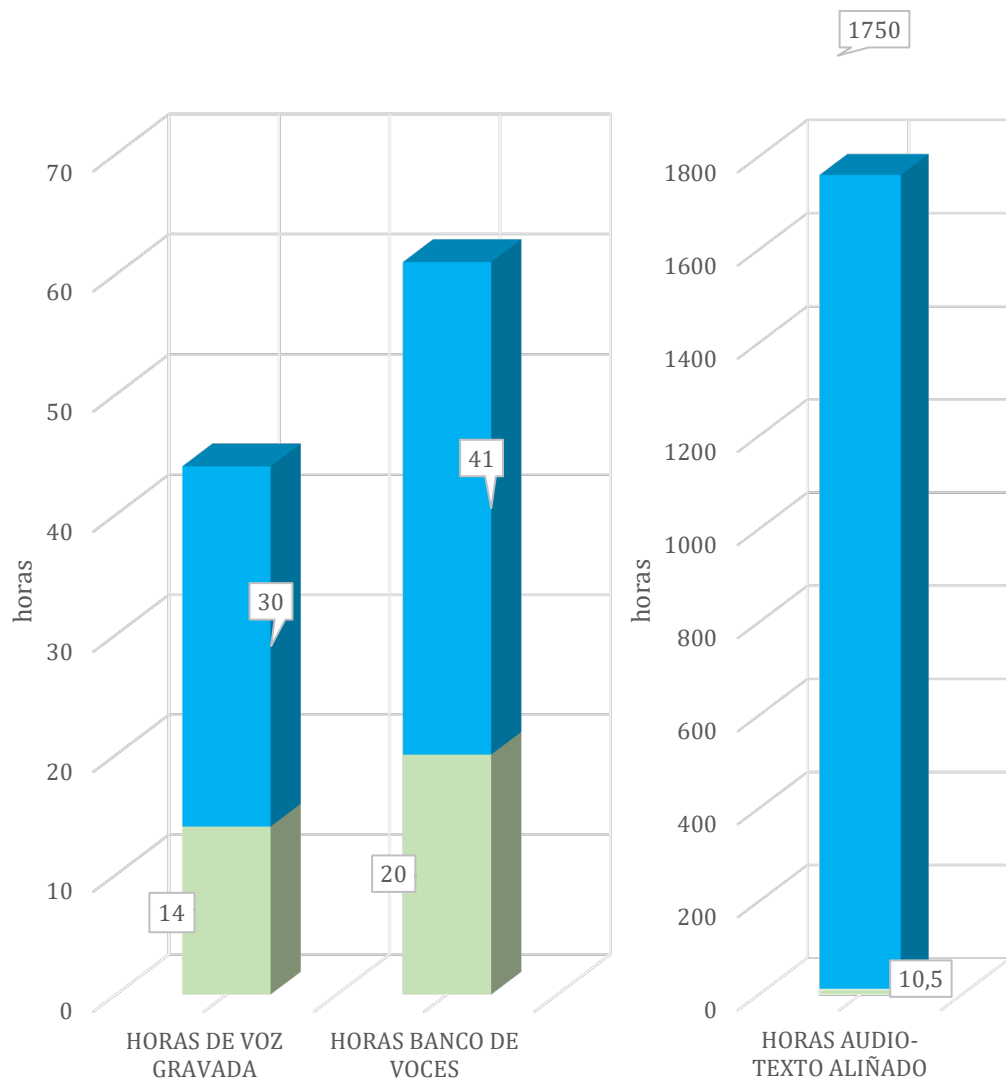
+ DE 40 ENTIDADES CONTACTADAS

17 ACORDOS DE CESIÓN asinados

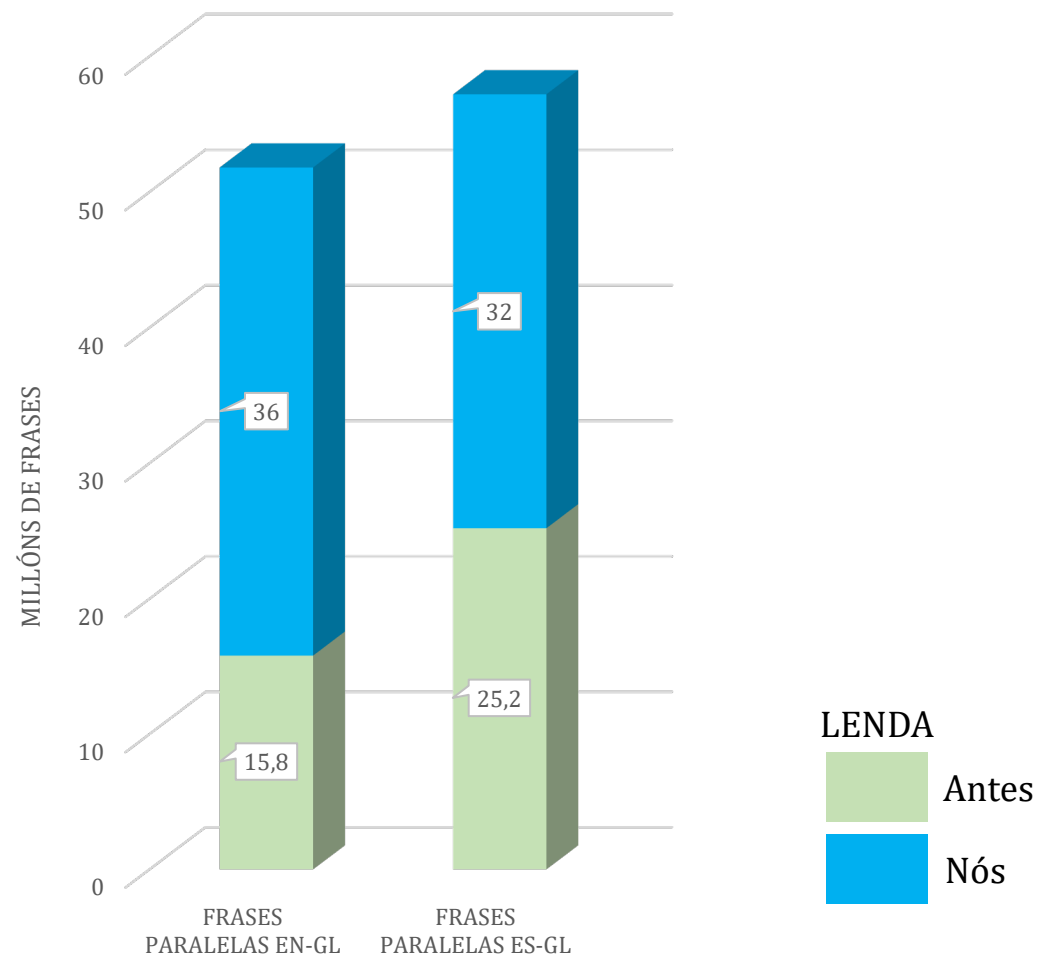


PARLAMENTO DE GALICIA | CONSELLO DA CULTURA GALEGA | RAG | CENTRO RAMÓN PIÑEIRO | CRTVG | DEPUTACIÓN DA CORUÑA | DEPUTACIÓN DE LUGO | EDITORIAL GALAXIA | EDITORIAL HUGIN E MUNIN | EDITORIAL LAIOVENTO | PRAZA PÚBLICA | SERMOS GALIZA | AGROCOMUNICACION | PASA NA COSTA | ATLÁNTICA DE INFORMACIÓN | IGADI | ASOCIACIÓN IRIMIA

VOZ



TRADUCCIÓN



LEND

- Antes
- Nós

RECOÑECEDOR AUTOMÁTICO DA FALA

MELLORAS E AVANCES

- Libre e de código aberto, **API DISPOÑIBLE**
- Primeiro recoñecedor neuronal en galego
- **PROPÓSITO XERAL**
- Modelo adestrado cun número reducido de datos (audio-texto aliñado)



Este é un sistema de recoñecemento de voz (ASR ou "Automatic Speech Recognition") en lingua galega baseado en redes neuronais artificiais. Teña en conta que as funcionalidades incluídas nesta páxina ofrécese unicamente con fins de demostración. Se ten algún comentario, suxestión ou detecta algún problema durante a demostración, póñase en [contacto](#) connosco.


Subir ficheiro

Comezar gravación

Transcribir

Modelo 0



0  Copiar

Borrar texto

SINTETIZADOR DE VOZ



Este é un sistema de conversión de texto a voz (TTS ou "Text-To-Speech") en lingua galega baseado en redes neuronais artificiais. Ten en conta que as funcionalidades incluídas nesta páxina ofrécense unicamente con fins de demostración. Se tes algún comentario, suxestión ou detectas algún problema durante a demostración, ponte en [contacto](#) connosco.

Celtia

Insira o texto

0/1000

Xerar voz

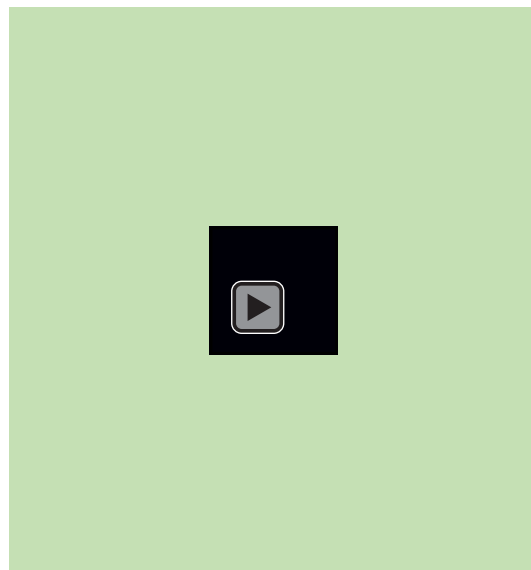
Borrar texto



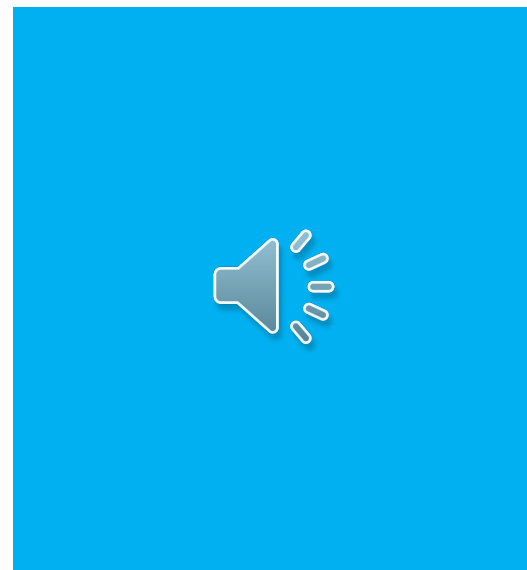
MELLORAS E AVANCES

- Primeiro modelo neuronal libre
- **A VOZ é LIBRE**
- Adestrado cun corpus de alta calidade
- Mellora da **CALIDADE DA VOZ**

CARMELA



CELTIA



LENDA

 Antes

 Nós

TRADUTOR NEURONAL

MELLORAS E AVANCES

- Primeiro **TRADUTOR NEURONAL** feito na Galiza
- Libre e de **CÓDIGO ABERTO**
- **CALIDADE** próxima a *Google translator*
- Neste momento xa é potencialmente **ÚTIL** para terceiros

Esta é un sistema de **tradución automática** de propósito xeral baseada en **redes neuronais artificiais** que está en fase de probas. Esta tecnoloxía ten a particularidade de depender da dispoñibilidade de grandes cantidades de datos, e que polo momento as traducións xeradas de forma automática non son totalmente fiables e precisan dunha revisión experta. É por iso que solicitamos a túa colaboración e convidámoste a que nos fagas chegar a túa opinión sobre a calidade das traducións premendo en "*axúdanos a mellorar*". Polo momento esta ferramenta só traduce de Castelán a Galego e de Inglés a Galego. Estamos traballando para engadir as outras direccións (Galego a Castelán e Galego a Inglés) así como para mellorar a calidade dos modelos.

Galego Español Inglés ⇌ Inglés Galego Español Transformer

Insira o texto ou URL que desexa traducir

0/1000

TRADUCIR

Limpar

Copiar

Axúdanos a mellorar

Advertencia

Os datos de uso do tradutor serán gardados e utilizados exclusivamente para a mellora do sistema. Ten en conta que as traducións xeradas de forma automática non son totalmente fiables e precisan dunha revisión experta.

Recursos accesibles e libres

The screenshot shows the Zenodo website interface. At the top, there is a search bar and navigation links for 'Upload' and 'Communities'. The main heading is 'Galician AI Language Resources'. Below this, there is a 'Recent uploads' section with a search bar and a 'New upload' button. Two datasets are listed: 'MeteoGalicia-GL Data-to-Text' (uploaded February 21, 2023) and 'Machine translation test suite English-Galician' (uploaded February 20, 2023). Each entry includes the dataset name, version, and a 'View' button. The 'MeteoGalicia-GL Data-to-Text' entry provides a brief description of the dataset and its origin. The 'Machine translation test suite English-Galician' entry describes the dataset's purpose for machine translation evaluation. On the right side, there is a community profile for 'Proxecto Nós' with a logo and a description of the community's focus on Galician AI language resources.

The screenshot shows the Hugging Face website interface. At the top, there is a search bar and navigation links for 'Models', 'Datasets', 'Spaces', 'Docs', 'Solutions', and 'Pricing'. The main heading is 'Proxecto Nós' with a logo and a 'Non-profit' badge. Below this, there is an 'Organization Card' for 'Proxecto Nós' featuring a video thumbnail and a list of 'Models'. The 'Models' section lists four datasets: 'proxectonos/NOS-MT-OpenNMT-es-g1', 'proxectonos/NOS-MT-OpenNMT-en-g1', 'proxectonos/data-to-text-g1', and 'proxectonos/wavvec2-large-x1sr-53-galicia...'. Each model entry includes the model name, version, and a 'private' status. The 'Organization Card' also includes a 'Research interests' section (currently empty) and a 'Team members' section with 12 members.

The screenshot shows the Proxecto Nós website interface. At the top, there is a search bar and navigation links for 'Product', 'Solutions', 'Open Source', and 'Pricing'. The main heading is 'Proxecto Nós' with a logo and a 'Non-profit' badge. Below this, there is a 'README.md' section with a link to the English text. The main content is a section titled 'Proxecto Nós: o galego na sociedade e na economía da intelixencia artificial'. This section includes a paragraph about the project's goals and a list of 'Repositories'. The 'Repositories' section lists three items: 'Corpus: corpora monolingües e multilingües de fala e texto para galego', 'Language-models: modelos lingüísticos a seren descargados para distintas aplicacións de procesamento da linguaxe natural', and 'Demos: páxinas web de demostración dos mellores modelos desenvolvidos polo equipo do Proxecto Nós'. Below this, there is a section titled 'Licenzas de corpus e modelos da linguaxe' which lists the licenses used for the corpus and models. The next section is 'Contribución ao proxecto - como podes envolverte na comunidade?' which provides information on how to contribute to the project. The final section is 'Para estar ao día do proxecto podes...' which lists ways to stay updated, such as visiting the website, emailing, or following on Twitter. The bottom section is 'Agradecementos' which acknowledges the funding sources for the project.

Vídeo de exemplo





PRÓXIMOS PASOS

PERTE DA LINGUA

2023-2025

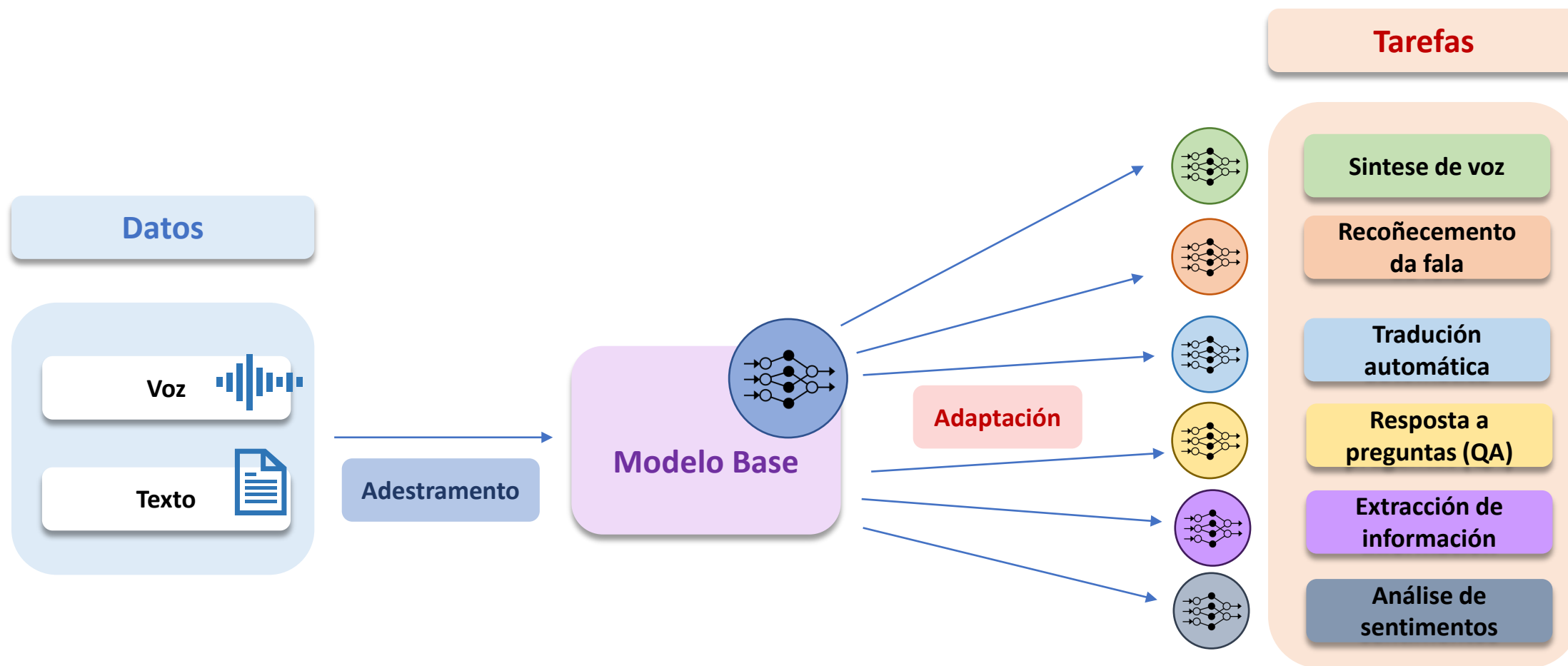
PROXECTO ILENIA

Proxecto coordinado polo BSC (participa USC, UPV, Universitat d'Alacant)

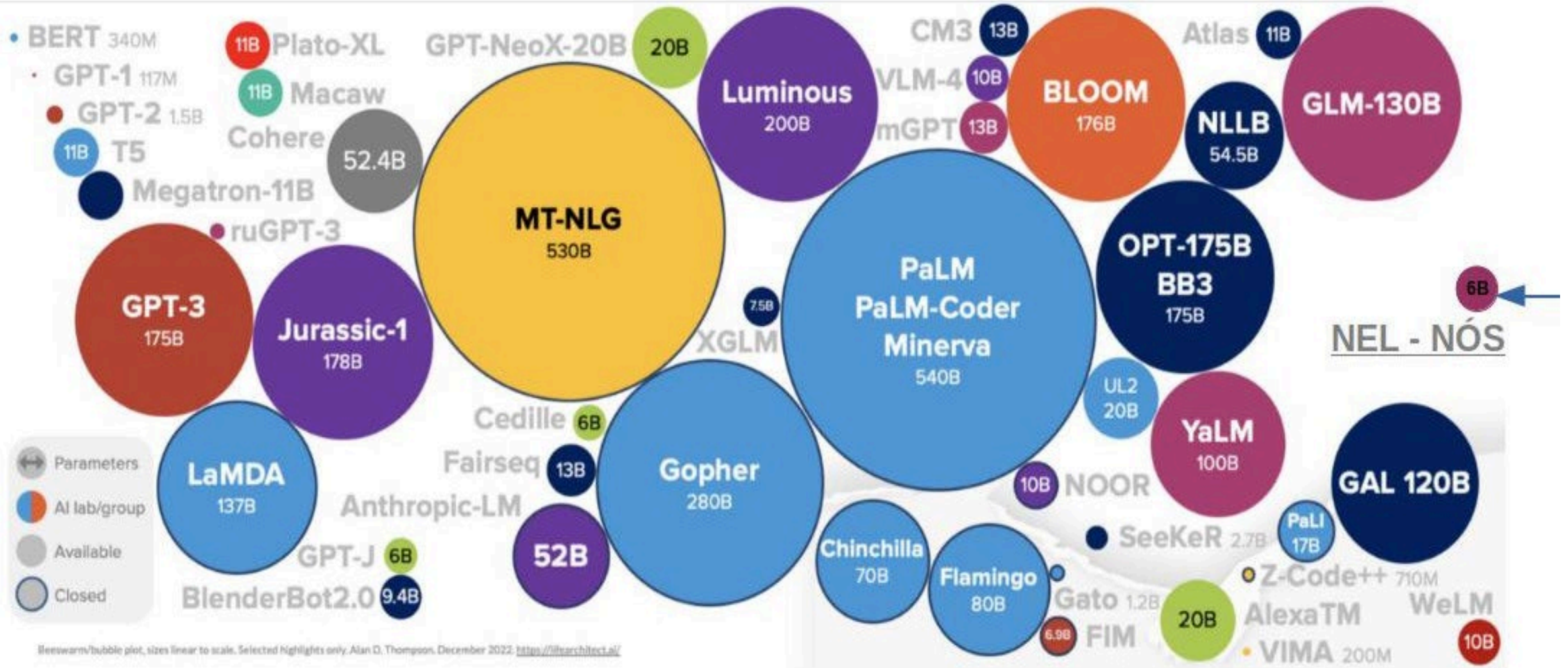
Obxectivos da USC

- Compilación de **grandes cantidades de corpus de galego** de calidade (acordos con fornecedores de recursos e campañas de recollida de audio para a elaboración de corpus de voz)
- Grandes **modelos multilingües** de texto, voz e tradución e modelos monolingües do galego
- Casos de uso

GRANDES MODELOS DE LINGUA



Modelos de lingua por tamaño





Usuarios do CESGA

Utilización do CESGA para o desenvolvemento de modelos de MT e GenerativeAI

- **Modelos de tradución automática EN-GL, ES-GL.**
- **Modelos de tradución automática multilingüe ES-GL-CA-EU**
- **Modelos de GenerativeAI (GPT) con modelo Nemo, Cerebras e Bloom.**

