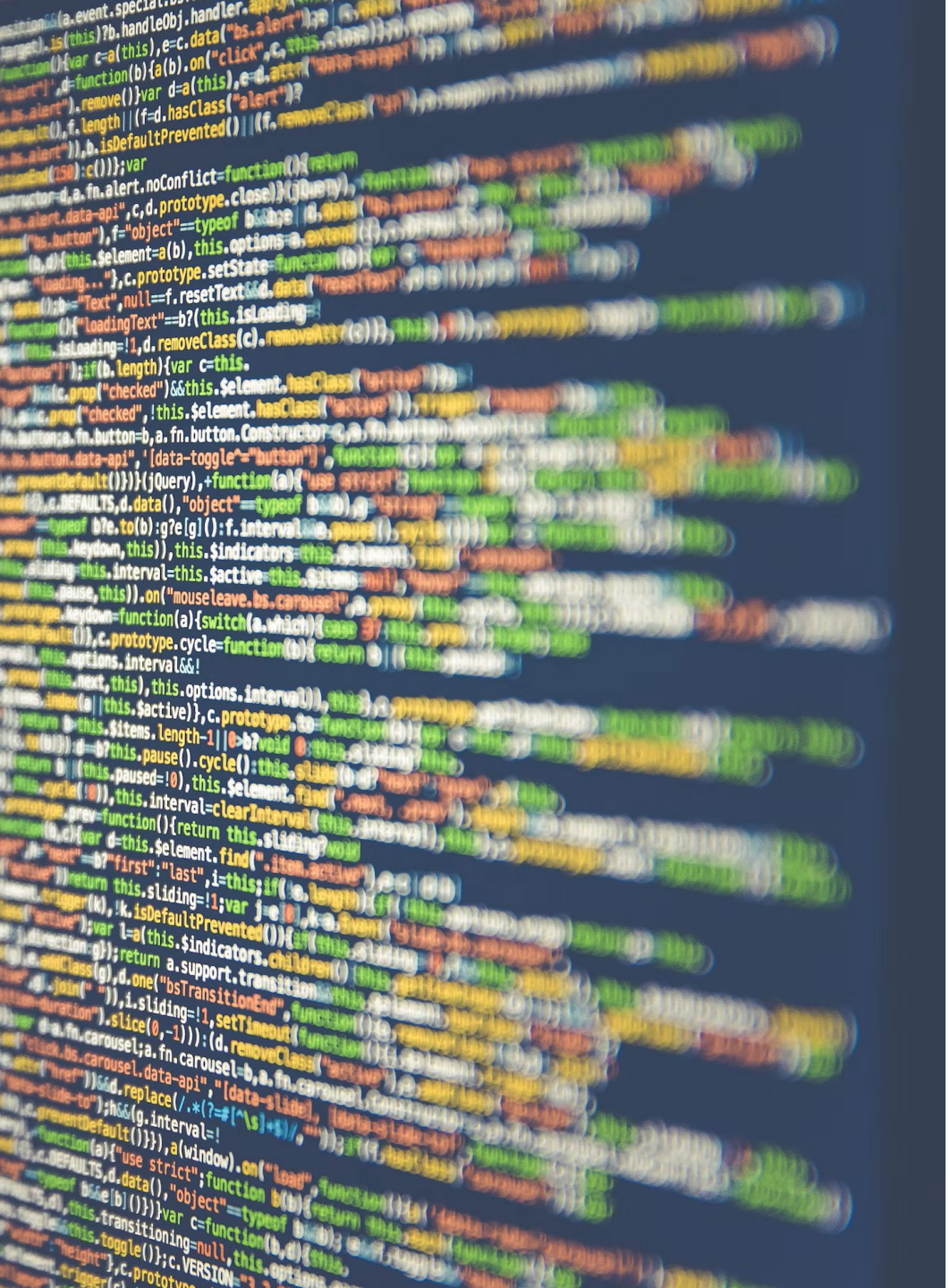
Proxecto Nós: o galego na vangarda das tecnoloxías intelixentes

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

XUNTA DE GALICIA | CONSELLERÍA DE CULTURA, EDUCACIÓN E UNIVERSIDADE
AXENCIA PARA A MODERNIZACIÓN TECNOLÓXICA DE GALICIA
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA INSTITUTO DA LINGUA GALEGA
CiTiUS Centro Singular de Investigación en Tecnoloxías Intelixentes

- 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

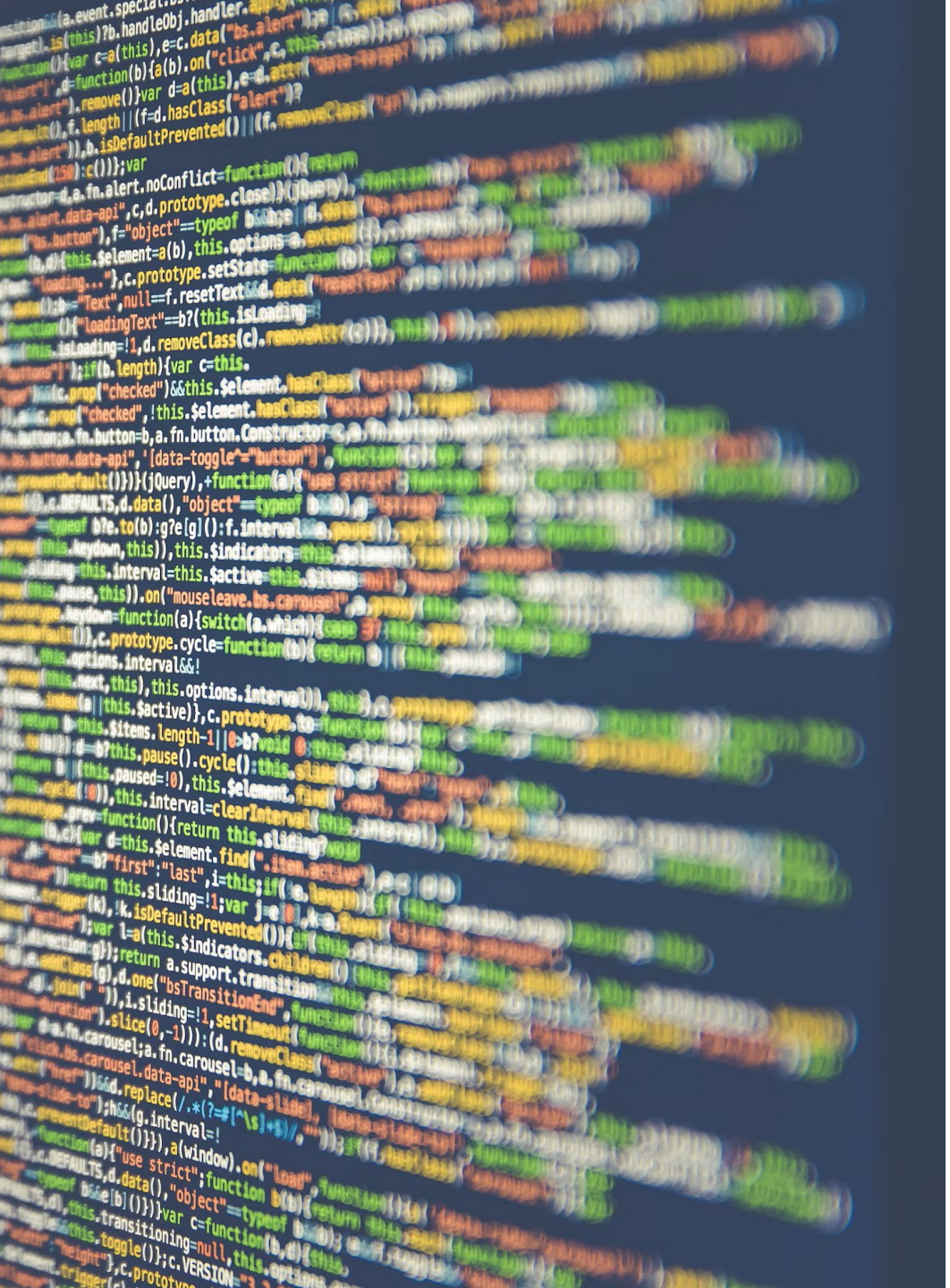
- O proxecto Nós, a través da USC recibe **2M€ do PERTE** Nova Economía da Lingua para o período **2023-2025**, co obxectivo de:
  - Aumentar a cantidade e calidade dos corpus actuais.
  - Elaborar **modelos** de voz e texto, tanto monolingües como **multilingües**.
  - Xerar datos anotados de calidade 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



nós

TRABALLO INICIAL

CONVENIO XUNTA / USC

2022

CONVENIO XUNTA / USC

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

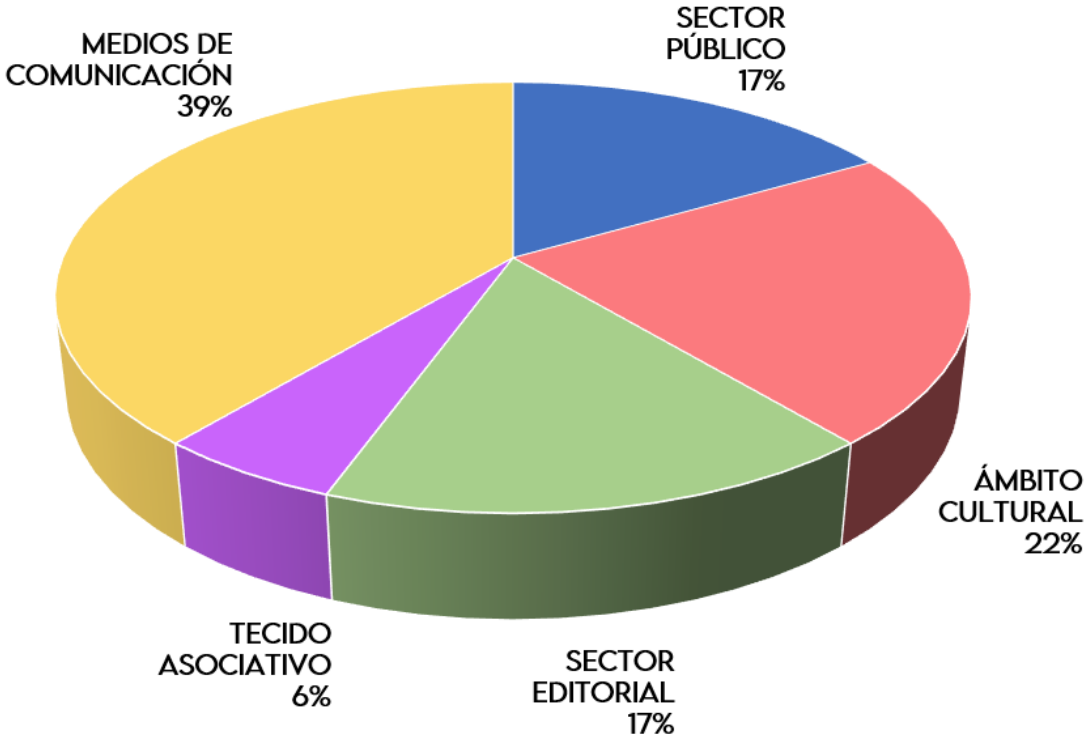
DATOS			
Corpus ASR (aliñado)	1750 h	Parlamento (2015-2022) + entrevistas e discursos	Lingua estándar (formal)
Corpus para ASR (datos brutos)	3700 h	TX 2019-2022 + AGO	Lingua estándar + popular
Corpus TTS	30 h	20.000 frases (extraídas)	Lectura
Corpus paralelo galego-español	30 millóns de frases	Parlamento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

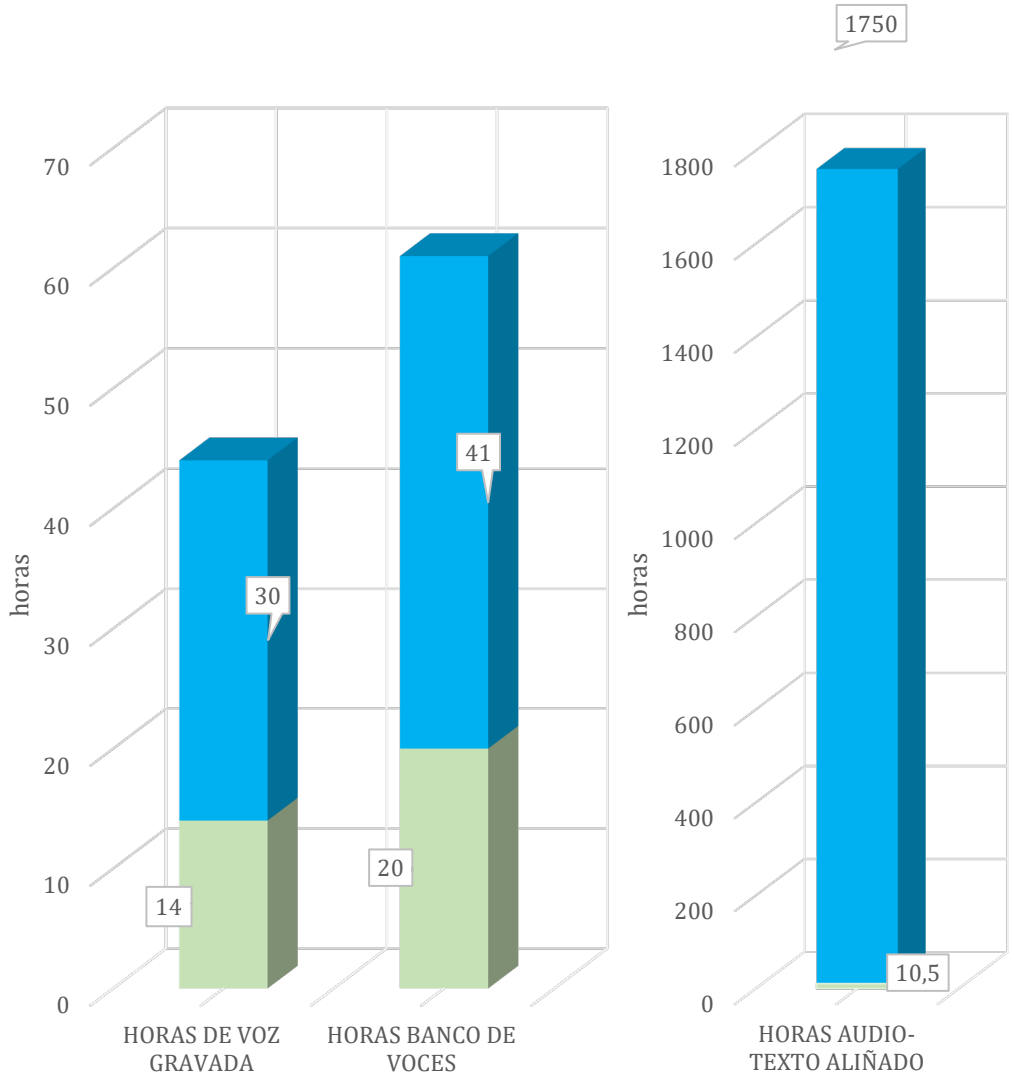
+ DE 40 ENTIDADES CONTACTADAS

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



TRADUCIÓN

LENDA

Antes

Nós

RECOÑECEDOR AUTOMATICO DA FALA

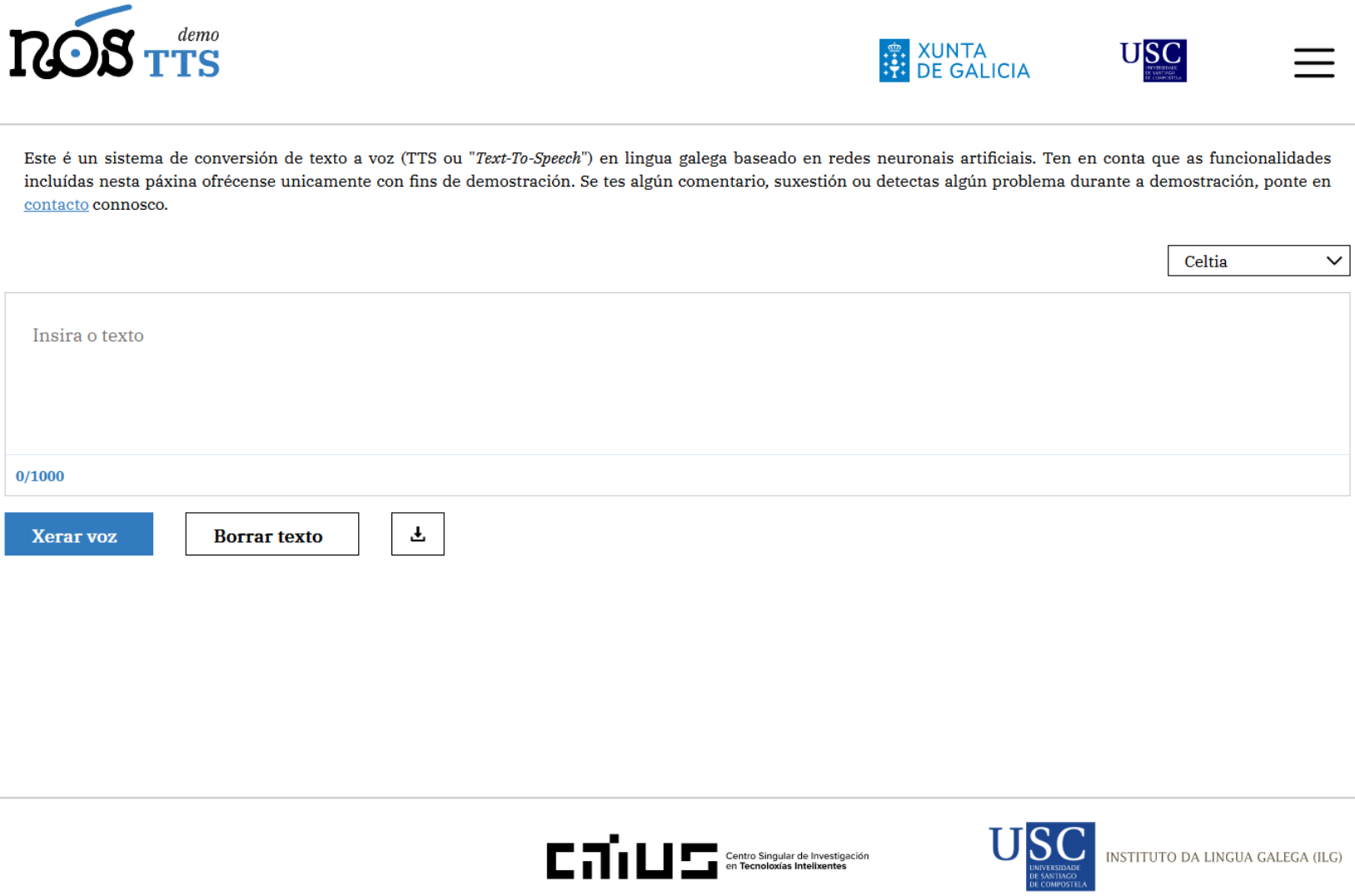
MELLORAS E AVANCES

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

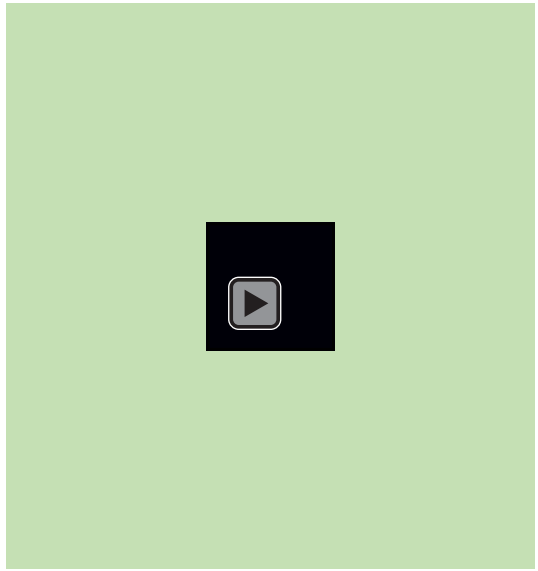
SINTETIZADOR DE VOZ

MELLORAS E AVANCES

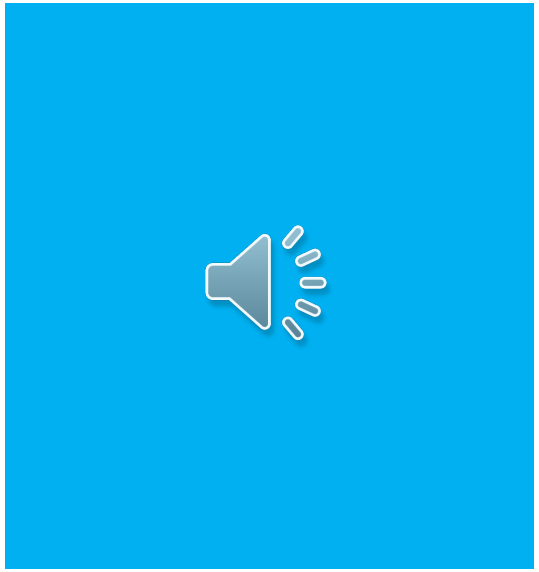
- Primero modelo neuronal libre
- A **VOZ** é **LIBRE**
- Adestrado cun corpus de alta calidade
- Mellora da **CALIDADE DA VOZ**



CARMELA



CELTIA



LENDA

Antes

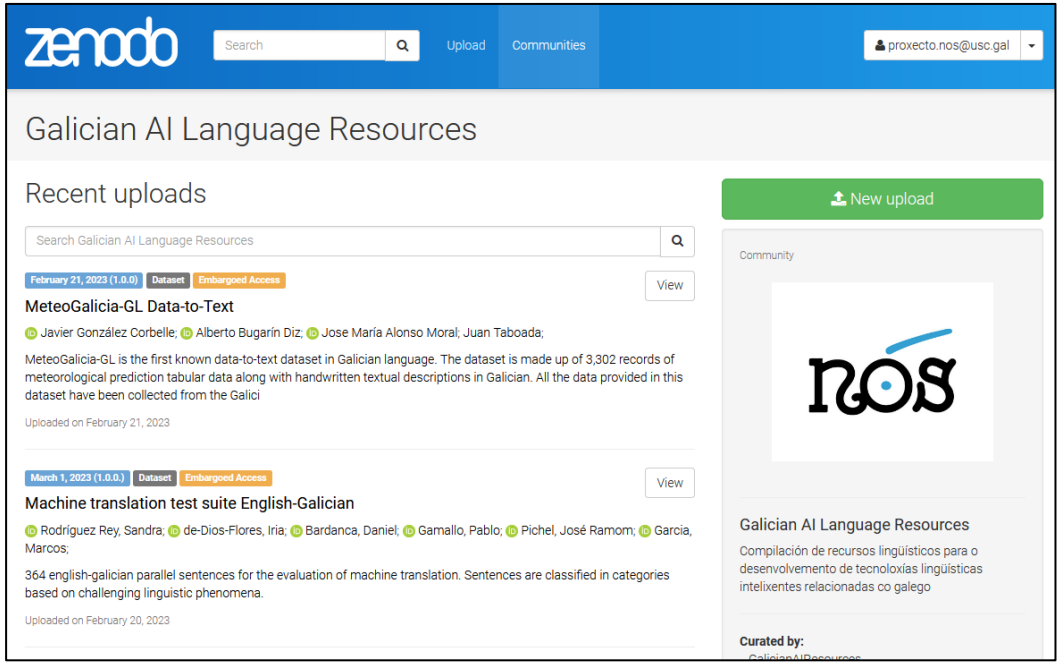
Nós

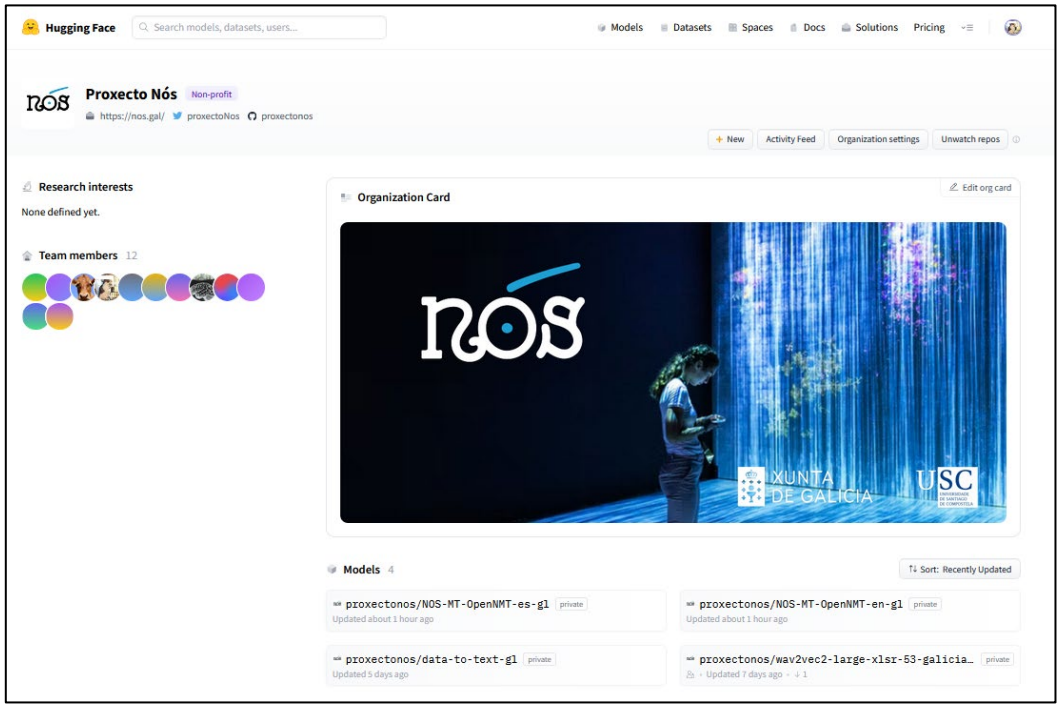
TRADUTOR NEURONAL

MELLORAS E AVANCES

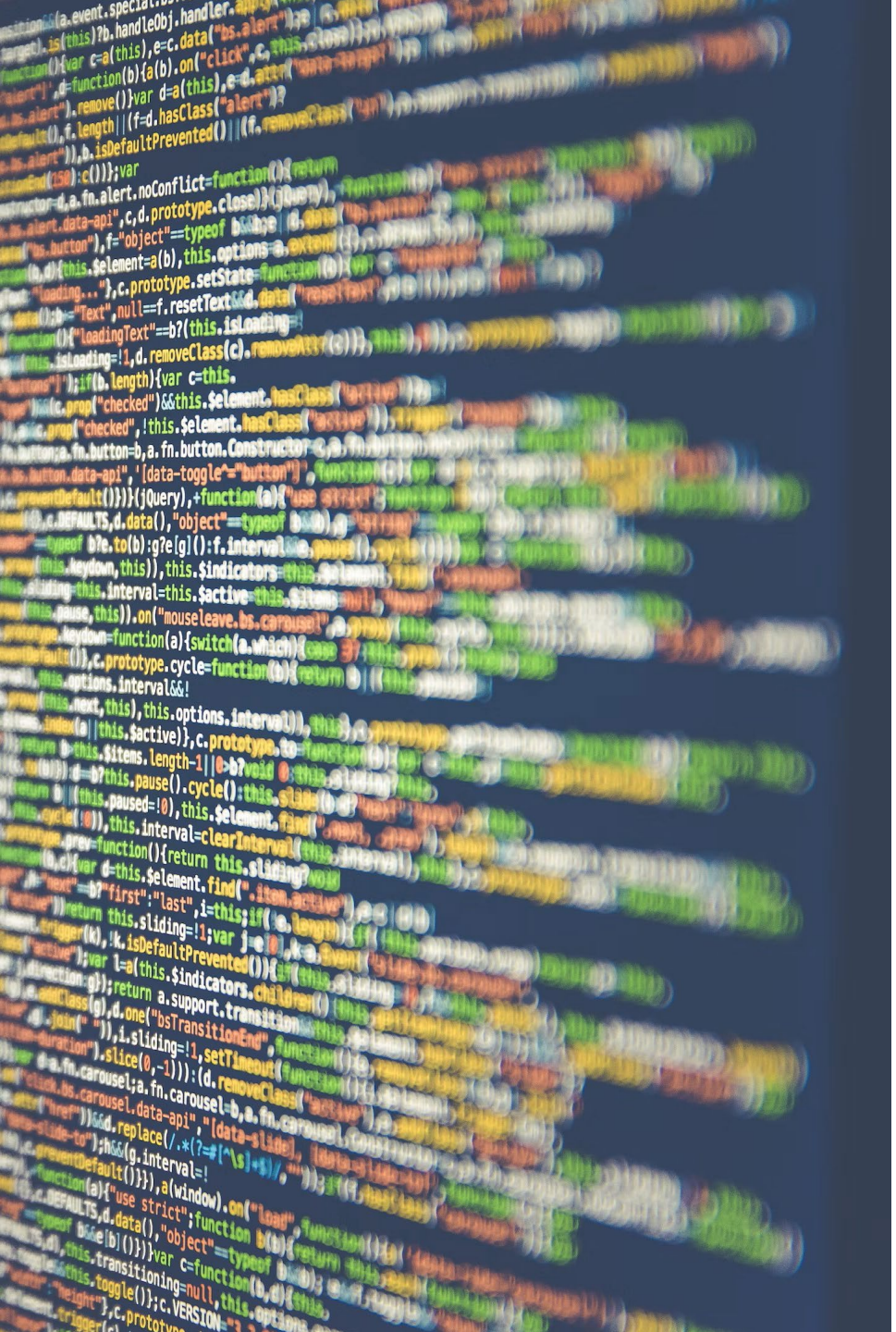
- Primeiro **TRADUTOR NEURONAL** feito na Galiza
- Libre e de **CÓDIGO ABERTO**
- **CALIDADE** próxima a *Google translator*
- Neste momento xa é potencialmente **ÚTIL** para terceiros

Recursos accesibles e libres





Vídeo de exemplo



PRÓXIMOS PASOS

PERTE DA LINGUA

2023-2025

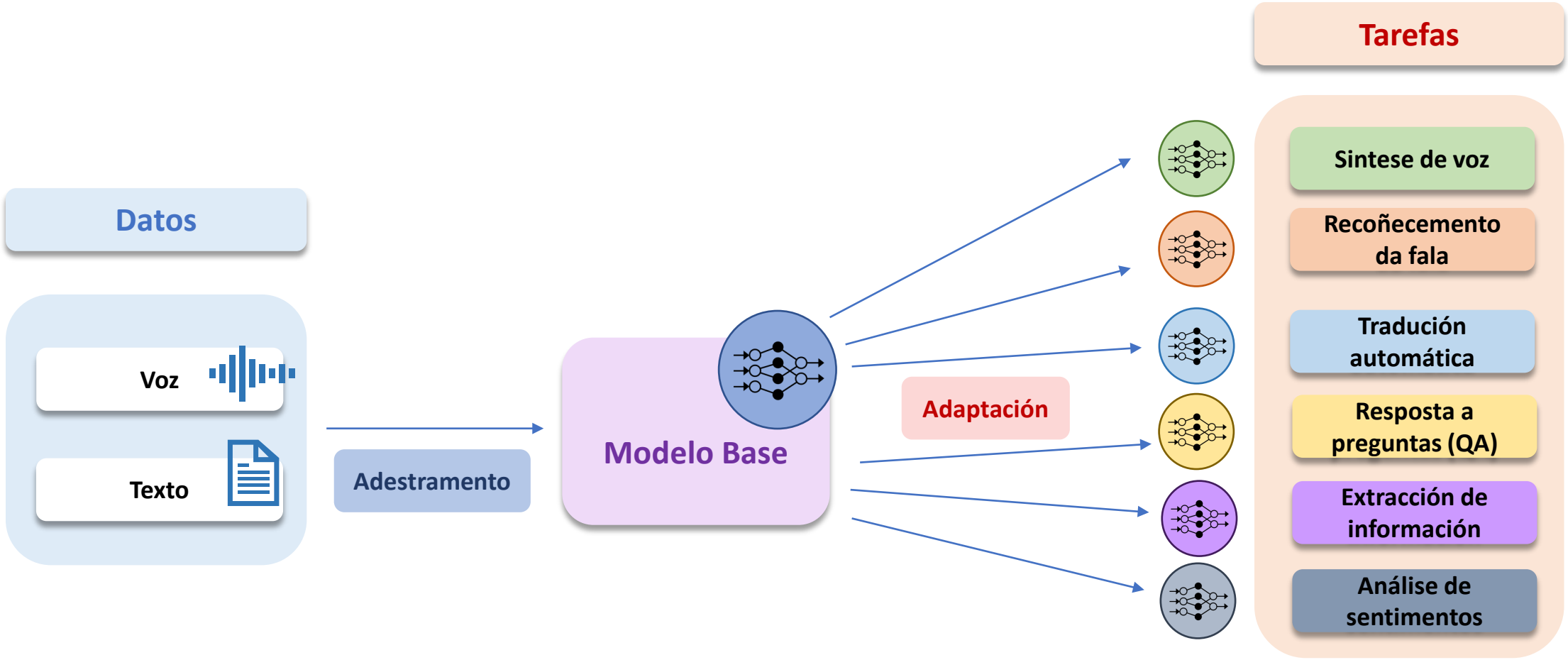
PROXECTO ILENIA

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

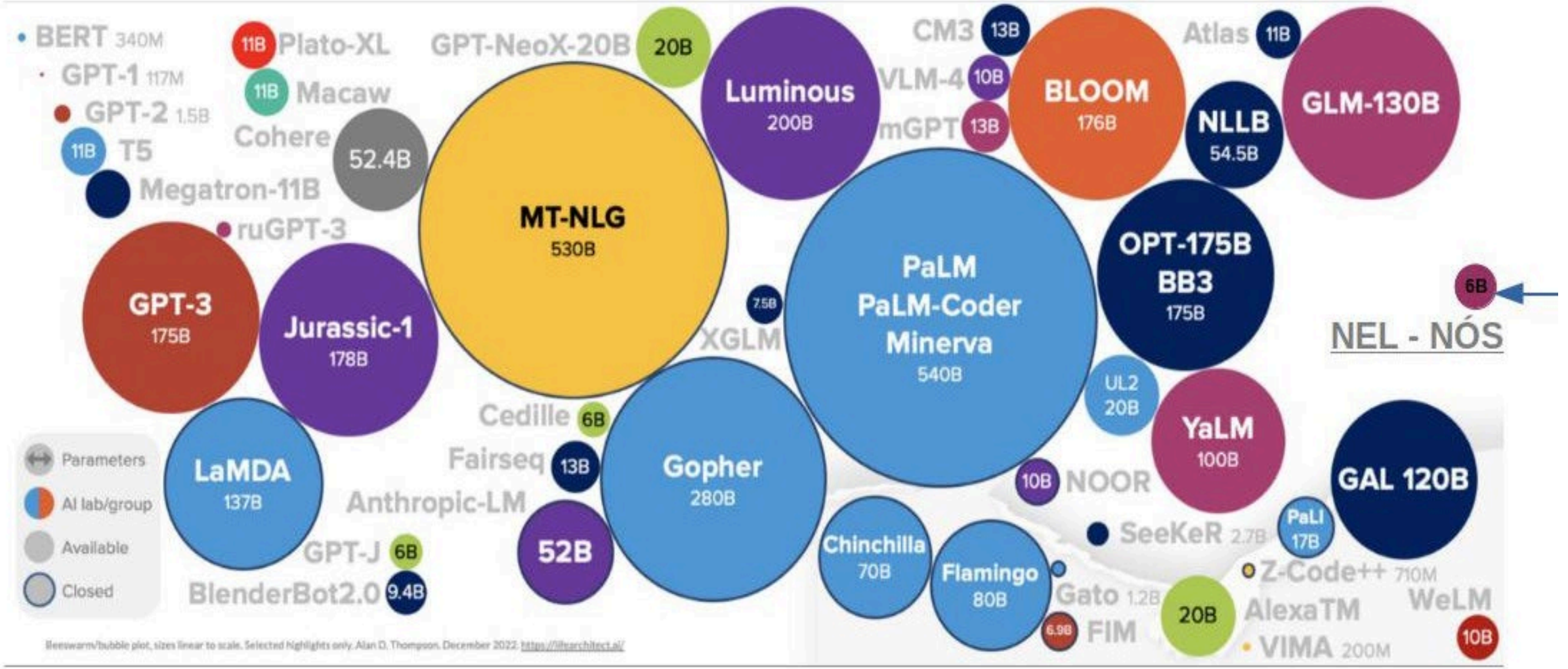
Obxectivos da USC

- 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.**