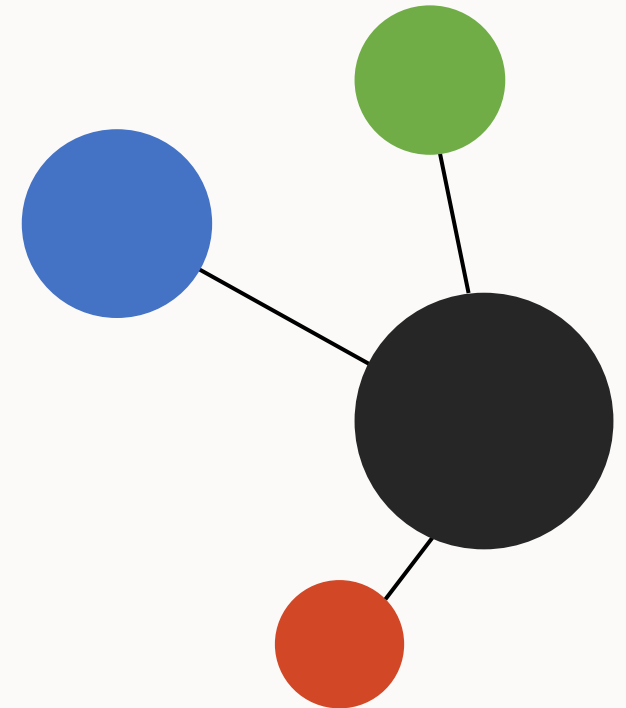


Ferramentas e recursos lingüísticos do Instituto da Lingua Galega

María Álvarez de la Granja

Elisa Fernández Rei

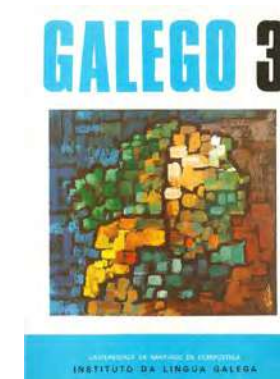
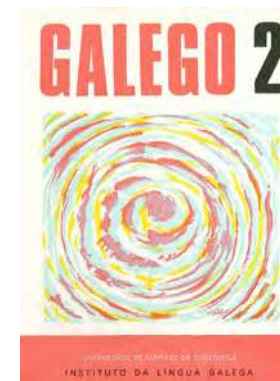
Instituto da Lingua Galega-Universidade de Santiago de Compostela



Instituto da Lingua Galega (ILG)

O ILG, fundado no ano 1971, foi o primeiro instituto de investigación da USC.

- ✓ Actividade investigadora nos eidos da lingüística e da filoloxía galegas.
- ✓ Resposta ás necesidades da sociedade galega.
 - Métodos de aprendizaxe do galego.
 - Código normativo.
- ✓ Desenvolvemento de recursos, ferramentas e aplicacións tecnolóxicas (de acceso libre).



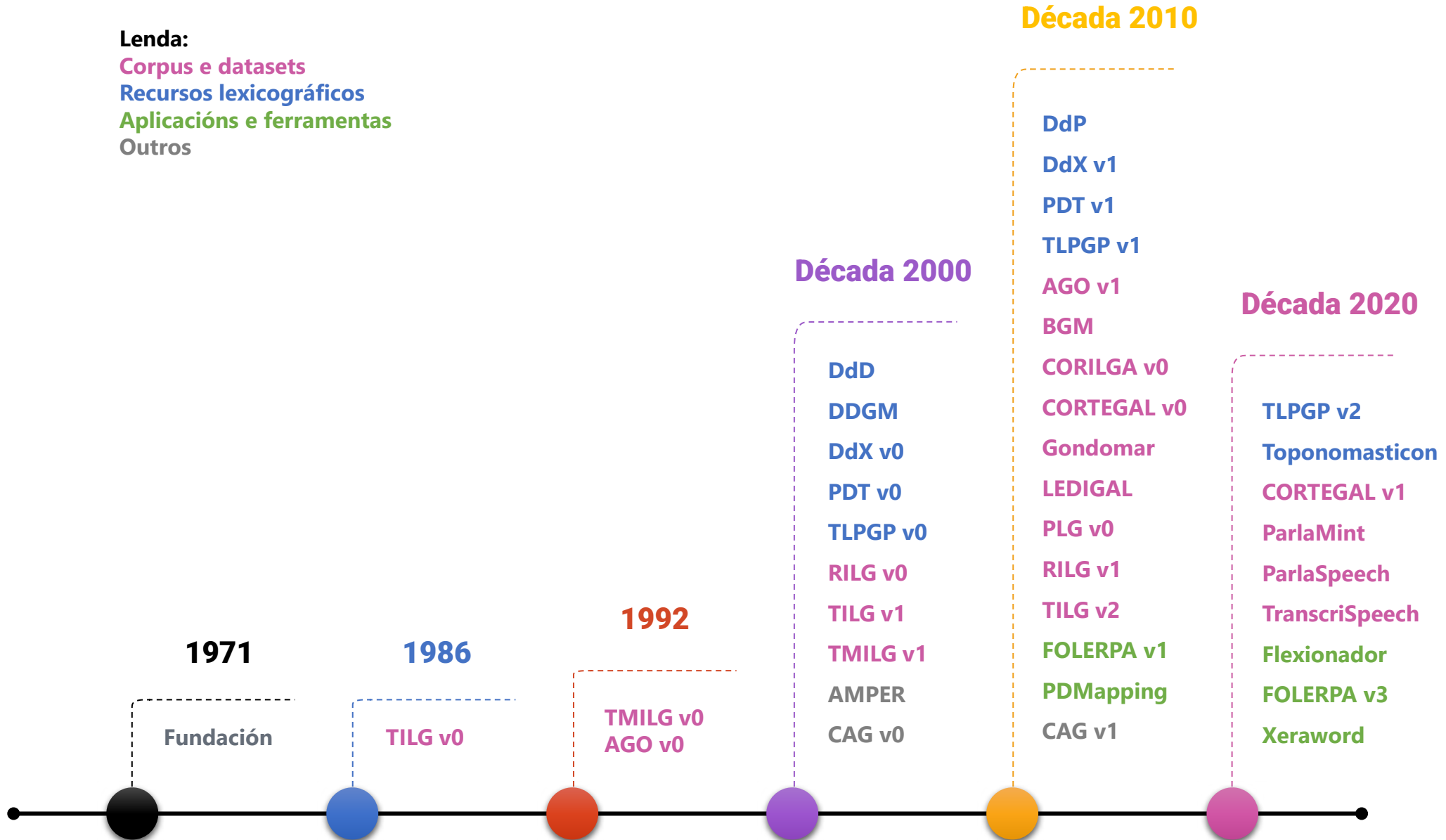
Lenda:

Corpus e datasets

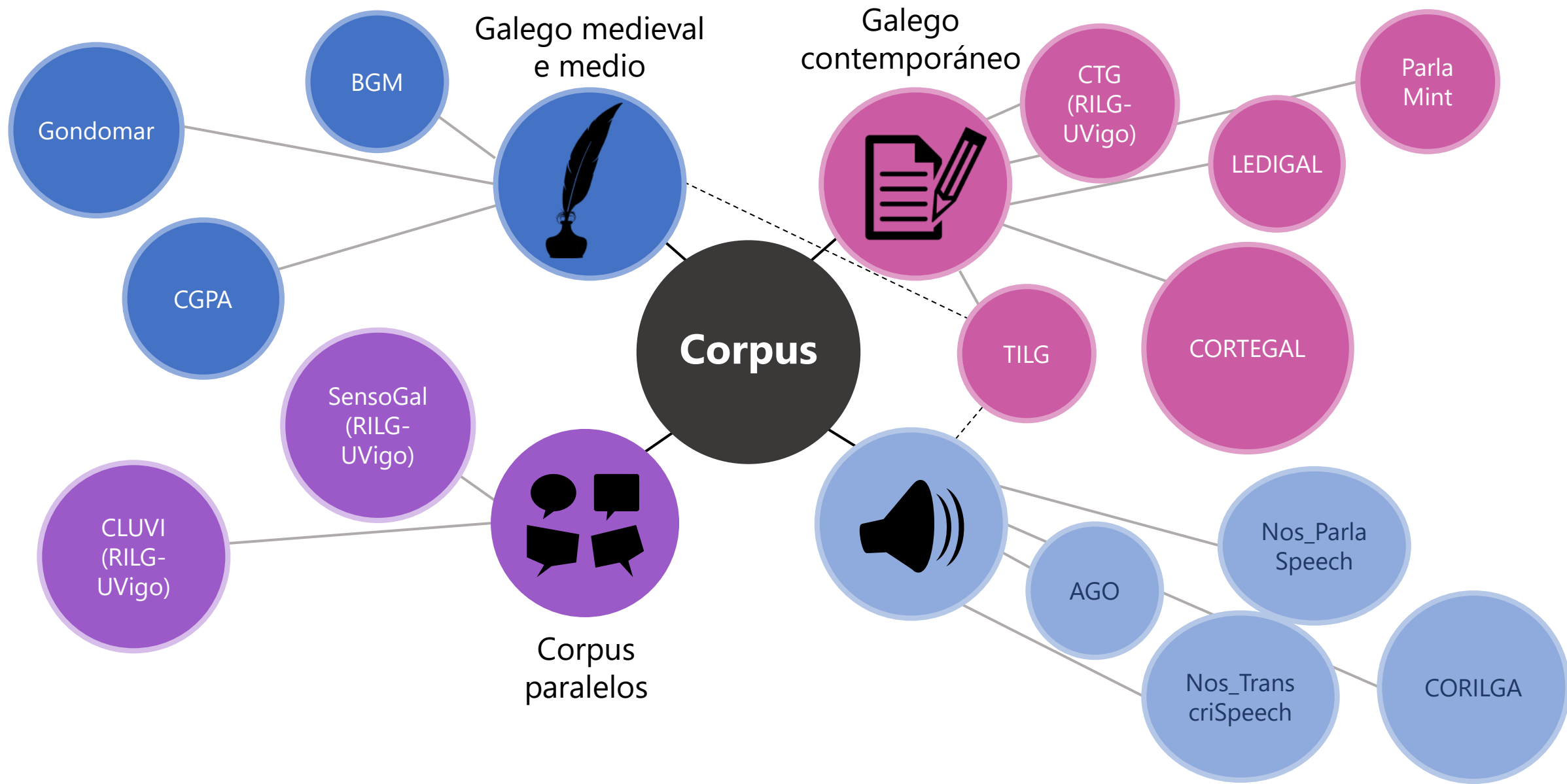
Recursos lexicográficos

Aplicacións e ferramentas

Outros



Corpus: tipos e períodos



Gondomar. Corpus dixital de textos galegos da Idade Moderna

- Corpus dirixido por Rosario Álvarez e Ernesto González Seoane e publicado en 2017
- Desenvolvido en TEITOK
- Produción escrita en lingua galega ao longo dos séculos XVI, XVII e XVIII
- 84 textos (1594-1791)
 - **Literarios:** poesía culta, poesía popular, poesía cancioneril, textos dramáticos e dialogados
 - **Non literarios:** correspondencia privada



Gondomar. Corpus dixital de textos galegos da Idade Moderna

- ✓ Visualización dos manuscritos orixinais e dos textos transcritos / editados
- ✓ Multiedición:
 - transcripción paleográfica ou moi conservadora
 - transcripción estendida
 - edición interpretativa
 - estandarización

Opcións de visualización

Texto:

Mostrar:

Anotación:

Eu beixo a vosa mercé as mans moitas veces por a memoria que tivo de acordarse de min; suplico a vosa mercé que non me olvide. A miña señora dona Costanza fago o mesmo, que é beixarlle as mans e que ningún no mundo é máis seu servidor ca eu, e como tal terei a moita dita se acorde de empregarme en cousas de seu servizo, con moita certeza que ninguén con tanta vontade coma eu a servirá, que en miña conciencia que, cando eu non tivera a obrigación que o mundo sabe, pola nobreza que en vosas mercés coñezo o fixera. A eses meus señores, seus fillos de vosa mercé e primos meus, beixo infinitas veces as mans, e Deus os faga en todo seus fillos de vosa mercé e de miña señora dona Costanza, a quen garde noso señor como eu, seu criado, desexo. Zamora, hoxe sábado. Seu sobriño de vosa mercé e servidor, don Xoán de Lanzós e de Andrade.

Manuscrito de Zamora, hoxe sábado.

Manuscrito de Zamora, hoxe sábado.

Gondomar. Corpus dixital de textos galegos da Idade Moderna

- Busca de documentos
- Busca nos textos
- Buscas CQL

Buscas no corpus

CQL Query: [Buscar](#) query builder | visualize | options

Caracteres especiais: ~u = ũ, ~e = ë, ~i = ĩ

Query Builder

| Busca de texto | | Busca de documentos | |
|---------------------------|--|---------------------|---|
| Transcripción provisional | igual a <input type="text"/> | Autor | <input type="text"/> |
| Estandarización | comeza por <input type="text"/> <input type="text"/> | Título | <input type="text"/> |
| Etiqueta morfosintáctica | igual a <input type="text"/> | Ano | <input type="text"/> - <input type="text"/> |
| Lema | igual a <input type="text"/> | | |

[Add token](#)

Caracteres especiais: ~u = ũ, ~e = ë, ~i = ĩ

[Buscar](#) cancel | Axuda

Buscas no corpus

CQL Query: [Buscar](#) query builder | visualize | options

Caracteres especiais: ~u = ũ, ~e = ë, ~i = ĩ

29 resultados • ipm: 475.71

Texto: [Transcripción](#) [Estandarización](#) [Etiquetas morfosintácticas](#) [Lemas](#)

Anotación: [Etiqueta morfosintáctica](#) [Lema](#)

| | | | |
|------------------|------------------------------|--------------------|-----------------------------|
| Mostrar contexto | conque ollos á | podeche | ver comprida na santa |
| Mostrar contexto | , è mais todos cantos | poderon | fazêlo, |
| Mostrar contexto | q̄ depedir. se | podese | Sahiu de |
| Mostrar contexto | . se podese. se | podese | . se podese/. |
| Mostrar contexto | , díf, (si | podese | /. Vm. |
| Mostrar contexto | ¡cantè quen | poidera | caber paga) ò |
| Mostrar contexto | pareçer, mais como | pudenVençer | mais! |
| Mostrar contexto | , fixen todo canto | puden | Eu só unha |
| Mostrar contexto | ; mais como | puden | a quen nin ber |
| Mostrar contexto | con farrápos vellos, non | puden | , ainda irey dando na |
| Mostrar contexto | Arçobispo Decer moi ben se | pudera | vencer à quen non ver |
| | | | entrare, ala mais adrento |
| | | | Cael non |

Gondomar. Corpus dixital de textos galegos da Idade Moderna

Aplicacións e utilidades

- Estudos lingüísticos: coñecemento do galego medio, análises diacrónicas
- Coñecemento e difusión de obras da Idade Moderna
- Aspectos socioculturais e sociolingüísticos: galego e castelán
- Normalización de textos históricos



Recursos lexicográficos



Tesouro do léxico patrimonial galego e portugués

- Recurso elaborado baixo a dirección de Rosario Álvarez e publicado en 2014 (actualización constante)
- Colaboración entre o ILG e universidades de Portugal e do Brasil
- Base de datos léxica que permite o acceso rápido e cómodo a información contida en traballos de léxico dialectal de galego, portugués de Portugal e portugués do Brasil
- Dicionarios e vocabularios dialectais, atlas lingüísticos, blogs con recolleitas de léxico, vocabulario recollido en centros de ensino... (desde 1917 á actualidade)

Lema Variante Buscar Lema... ?
 Sensible a diacríticos Sensible a maiúsculas

Busca avanzada ▼
Busca ampliada ▼

O *Tesouro do léxico patrimonial galego e portugués* é unha base de datos léxica que permite o acceso rápido e cómodo a información contida en traballos de léxico dialectal de galego, portugués de Portugal e portugués do Brasil.

O *Tesouro* permite consultar de xeito simple a información completa contida en obras lexicográficas e dialectais que non sempre son de fácil acceso para os investigadores. A información das fontes orixinais ofrécese completa e organizada a partir de variantes, lemas, clasificación semántica e localización. A información textual complementase cunha representación cartográfica da distribución xeográfica das formas.

O proxecto do *Tesouro* iniciouse en 2009 promovido polo Instituto da Lingua Galega (USC). Na actualidade colaboran nel 20 universidades de Galicia, Brasil e Portugal.

As buscas na base poden realizarse a partir das formas léxicas (variantes ou lemas), da localización (Galicia, Portugal e Brasil) ou do campo semántico co que se asocian.

O *Tesouro* continúa arrequentándose coa participación de todos os colaboradores. Desde 2014 a base de datos do *Tesouro* é accesible de maneira libre e gratuíta.



Datos xerais

Obras introducidas: 296 (Brasil 48, Galicia 155, Portugal 81, Galicia e Portugal 10, Galicia, Portugal, Brasil 2)
Número total de rexistros: 308255
Número de lemas: 74732
Número de variantes: 122490

Tesouro do léxico patrimonial galego e portugués

- Lematización das variantes recollidas nas obras
- Clasificación semántica
- Localización xeográfica



Lema Variante **toupa**

Sensible a diacríticos Sensible a maiúsculas

Busca avanzada ▼
Busca ampliada ▼

Localización (2)

- Galicia (41)
- Portugal (4)

Campo semántico (4)

- 5.6 - Animais (enfermidades e lesións) (3)
- 6.1 - Animais silvestres (denominacións) (38)
- 6.2 - Animais silvestres (partes, costumes...) (2)
- 7.1 - Ser humano

Lemas (2)

- toupa (pt-eu) (4)
- toupa (gl) (41)

Variantes (12)

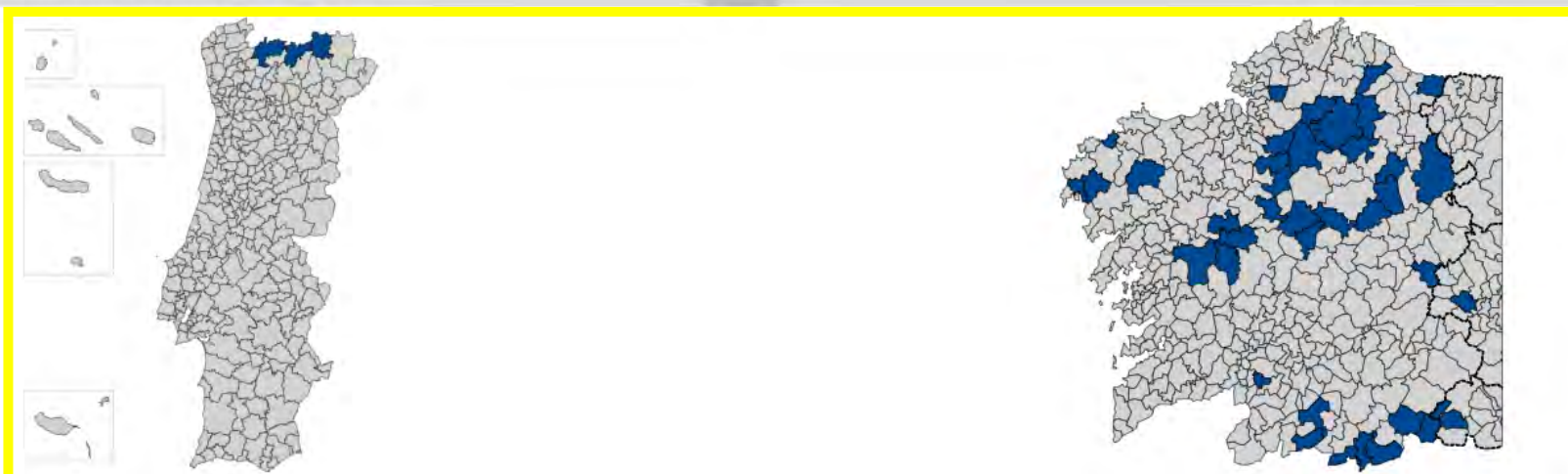
- toupa (pt-eu) (4)
- teupa (gl) (6)
- tiopa (gl) (3)
- topia (gl) (4)
- toupa (gl) (15)
- teipa (gl) (5)
- teipe (gl) (3)
- calandeiras de teipe

Categorías (1)

- substantivo feminino (45)

Resultados atopados (45)

| | | |
|------------------------|--|--------|
| teida <i>tl</i> | <i>lipa</i> Topo. Baamonde 1977:70. Vid. teinoeira . LEMA: toupa sf. | GL 6.1 |
| teipa <i>t</i> | <i>lipa</i> Mamífero de fucíño afiado, ollos pequenos, moi mouro; fai canalóns por debaixo da terra. {esp.} Topo. Pena 1970:18. LEMA: toupa sf. | GL 6.1 |
| teipa <i>s</i> | {Aldixe, Carballido, Castromaior, Corvelle, Insua, Oleiros, Árbol, Xoibán} Toupa. V. teupa . Regueira 1989. LEMA: toupa sf. | GL 6.1 |
| teipa <i>R</i> | elo 1971. Vid. teipe . LEMA: toupa sf. | GL 6.1 |
| teipa <i>t</i> | <i>lipa</i> Topo. Rodríguez 1971:49. LEMA: toupa sf. | GL 6.1 |
| teipe <i>s</i> | sf. Bicho sin ollos, negro, que fai galerías debaixo da terra. Castro 1986:65. LEMA: toupa sf. | GL 6.1 |
| teipe <i>s</i> | sf. Teipa . Rielo 1971:328. Vid. teipa . LEMA: toupa sf. | GL 6.1 |
| teipe <i>T</i> | Topo. Mamífero insectívoro, pequeno, negruzco, con hocico afilado que vive en galerías subterráneas, que él mesmo cava con las uñas. VázquezF 1971:32. LEMA: toupa sf. | GL 6.1 |
| téipoa <i>f</i> | {Rioaveso (San Xurxo)} V. teipa . Regueira 1989. LEMA: toupa sf. | GL 6.1 |
| teupa <i>T</i> | Topo. <i>Talpa occidentalis</i> , Cabrera Couceiro 1967:35. LEMA: toupa sf. | GL 6.1 |



Lema Variante **Definicións** Exemplos cereal ?
 Orde personalizada ?

Busca simple ▲
 Busca avanzada ▼

Localización (3)

- Galicia (312)
- Brasil (2)
- Portugal (283)

Campo semántico (24)

- 3.1 - Terras de cultivo e cerras (14)
- 3.2 - Labores agrícolas (51)
- 3.2.1 - Cereais e herba (300)
- 3.2.2 - Vide e viño (1)

Lemas (582)

- a chapodo (gl) (1)
- a grade (gl) (1)
- abrollar (gl) (1)
- acambolhado (pt-eu) (1)
- acarreja (pt-eu) (1)
- acarrear (pt-eu) (1)
- aconchegar (gl) (1)

Variantes (418)

- a chapodo (gl) (1)
- abea (gl) (1)
- aborcállarse (gl) (1)
- abrollar (gl) (1)
- acambolhado (pt-eu) (1)
- acarreja (pt-eu) (1)
- acarrear (pt-eu) (1)

Resultados atopados (597)

- a chapodo** loc. Modalidade de sega que consiste en corta-lo cereal dándolle golpes co fouciño LópezC 1990:94. LEMA: [a chapodo loc adv.](#) GL 3.2.1
- sementar a grade** Siembra empleada para los cereales. Lois 1969:123. Vid. [sementar](#). LEMA: [a grade loc adv.](#) GL 3.2
- abrollar** Dícese del cereal cuando está asomando todavía en la tierra. Taboada 1971:70. LEMA: [abrollar v.](#) GL 3.2.1
- acambolhado** adj. Diz-se do cereal, nas searas, quando, de forte, ajoelha um sobre o outro (Mogadouro e Lagoaça – AM). Observação: AH grafa acambulhado. BarrosV 2006:18. LEMA: [acambolhado a.](#) PT-EU 3.2.1
- acarreja** Transporte do cereal para a eira. SousaS et al. 2019. LEMA: [acarreja sf.](#) PT-EU 3.2.1
- acarrear** v.tr. Transportar o cereal para a eira, carrear (Fornos – FEC). BarrosV 2006:19. Vid. [carrear](#). LEMA: [acarrear v.](#) PT-EU 3.2.1
- aconchegar** [aconche\[th\]ar](#) En los cultivos de cereales o patatas, pasar un arado entre surco y surco para acercar tierra al pie de la planta. También puede hacerse con una azada o incluso a mano. Gil 1969:72. LEMA: [aconchegar v.](#) GL 3.2
- acuña** Desenvolverse defectuosamente los cereales por estar demasiado juntos. Couceiro 1967:104. LEMA: [acuña v.](#) GL 3.2
- adelha** Acóde-se à moega por escada e lança-se nela o cereal que sai para a adalha e desta para o olho da mó (...). Guerreiro 1982:130. Vid. [moega](#). LEMA: [adalha sf.](#) PT-EU 3.1.1

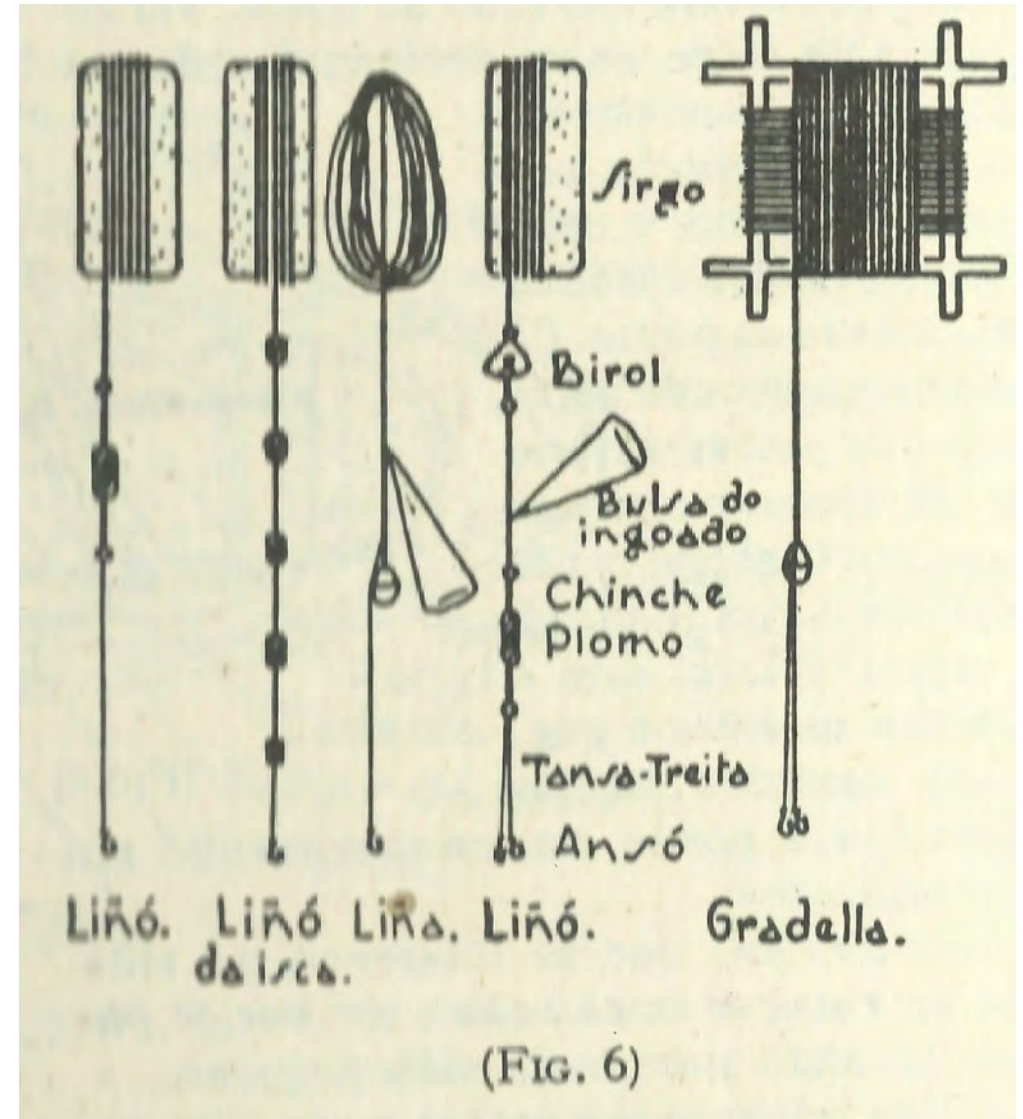
Tesouro do léxico patrimonial galego e portugués

- Busca por lemas
- Busca por variantes
- Busca por palabras das definicións e dos exemplos
- Busca por clases semánticas
- Busca por localizacións

Tesouro do léxico patrimonial galego e portugués

- **Aplicacións e utilidades**

- Xeolingüística
- Historia da lingua
- Estudos etnográficos
- Estudos contrastivos entre o galego e o portugués
- Consulta de obras de difícil acceso ou mesmo inéditas
- Coñecemento, visibilización e posta en valor do léxico dialectal

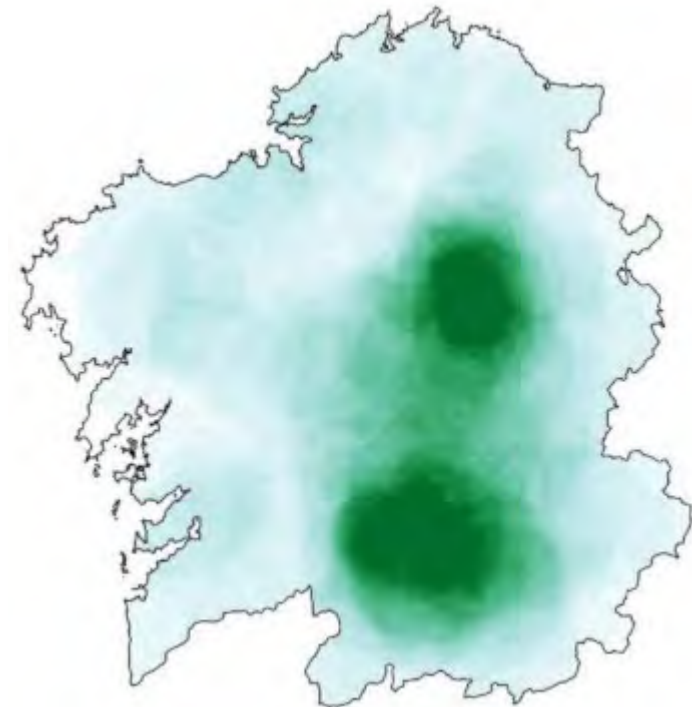
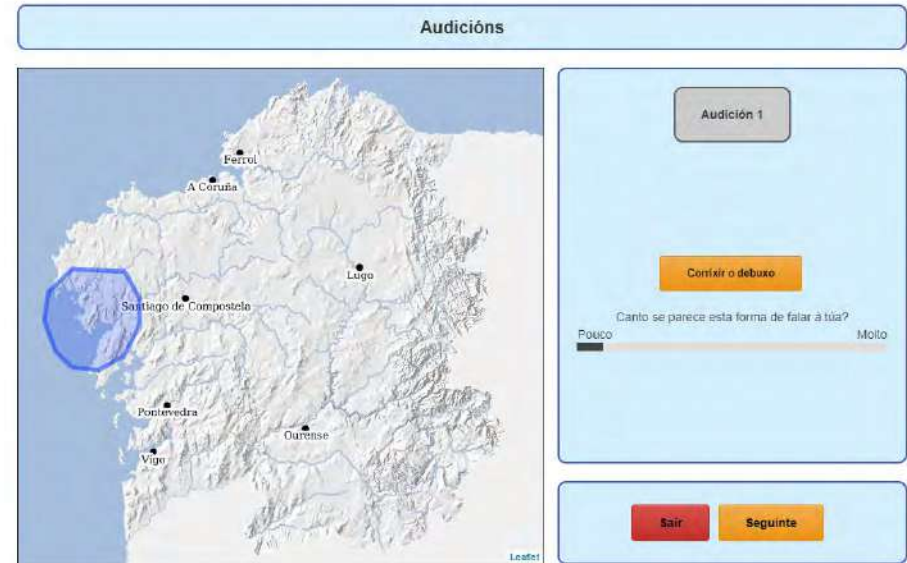


Aplicaciones e herramientas



PDMapping

- Permite realizar experimentos de Dialectoloxía Perceptiva (recollida e procesado de datos)
- Almacena datos en formato estándar (R e QGIS)
- Procedemento:
 - Variables: o/a informante fornece información sobre idade, educación, perfil lingüístico...
 - Test: escoita un estímulo e asígnalle unha área no mapa
 - Percepcións e actitudes: responde un cuestionario
- Código aberto (dispoñible en GitHub)




Colaboración co CiTIUS

proyecto_ **NÓS**

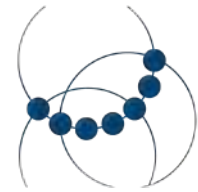
| MODELOS | | | |
|---|----------------------|--------------|---------------|
| Modelo | Utilidade | Arquitectura | Dispoñible en |
| proxectonos/Nos_TTS-celtia-vits-graphemes | TTS | VITS | HuggingFace |
| proxectonos/Nos_TTS-sabela-vits-phonemes | TTS | VITS | HuggingFace |
| proxectonos/Nos_ASR-wav2vec2-large-xlsr-53-gl-with-lm | ASR | wav2vec2 | HuggingFace |
| proxectonos/Nos_D2T-gl | Text2Text 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 |

| 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  ↗ |
|--|---|---|

Spanish CLARIN Centre-K

CLARIN
K-centre



- Centros: IXA-UPV e ILG-USC
- Obxectivo:
 - ofrecer os coñecementos e a experiencia dos grupos que o compoñen no emprego da tecnoloxía para a investigación en humanidades e ciencias sociais

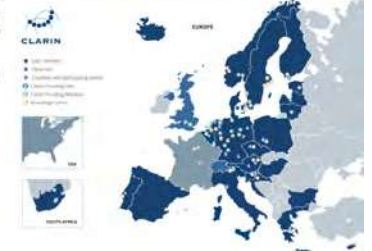


Bienvenido a CLARIN Centre-K

CLARIN Centro-K-Español forma parte de la infraestructura europea CLARIN y tiene como objetivo ofrecer los conocimientos y la experiencia de los tres grupos que inicialmente lo componen en la utilización de tecnología para la investigación en humanidades y ciencias sociales. Somos el primer Centro CLARIN Knowledge y la representación de la infraestructura CLARIN en España.

Como Centro-K, se reúnen las fortalezas de los tres centros participantes con el fin de ofrecer servicios a la comunidad de forma unificada y organizada. Los servicios que el CLARIN Centro-K-español ofrece son:

- Consultoría virtual para asesorar y responder dudas sobre cuestiones prácticas relacionadas con estándares, uso de herramientas de procesamiento y acceso a recursos lingüísticos. El centro ofrece contacto por correo electrónico con garantía de respuesta en 24 horas. Se trata de un servicio de orientación personalizada y que también pone al servicio de los investigadores documentación y casos de uso conocidos en los que encontrar buenas prácticas.
- Soporte para auto-aprendizaje con recursos especializados; catálogos especializados donde encontrar información sobre prácticas actuales y acceso directo a herramientas, en aplicaciones web, que facilitan el uso de las tecnologías existentes, videotutoriales, MSOCs, etc.

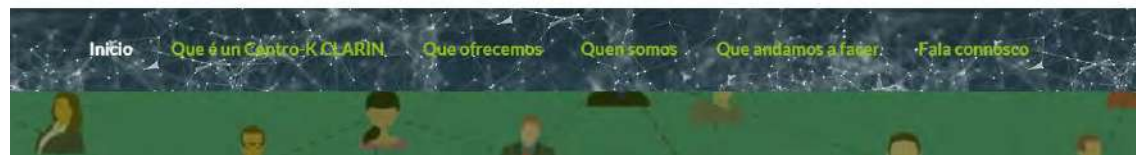


CLARIN
K CENTRE



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA
INSTITUTO DA LINGUA GALEGA

EN



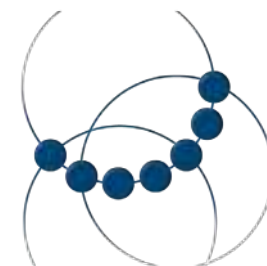
Inicio Nuestros productos Videos Actividades Servicios Contacto

CLARIN
K-centre

Servizos

- Consultoría virtual para asesorar e responder dúbidas sobre cuestións prácticas relacionadas con estándares, uso de ferramentas de procesamento e acceso a recursos lingüísticos.
- Soporte para autoaprendizaxe.
- Organización de programas de ensino e formación para investigadores/as, estudantes, proxectos ou grupos de interese.
- Servizos tecnolóxicos e de xestión e planificación de proxectos personalizados en función das necesidades dos/as solicitantes.

CLARIN
K-centre



Desafíos e oportunidades

- Actualización e mantemento
- Visibilización
- Mellora dos servizos a investigadores/as



Incorporación de recursos e herramientas ás infraestructuras europeas



- CLARIN Virtual Language Observatory
<https://vlo.clarin.eu/?1>



- Language Resource Switchboard
<<https://switchboard.clarin.eu/tools>>



Social Sciences & Humanities Open Marketplace
<https://marketplace.sshopencloud.eu/>

Bibliografía

- Alonso Pintos, Serafín (2017): "Investigar, elaborar, divulgar. O Instituto da Lingua da Universidade de Santiago de Compostela, *LaborHistórico*, 3/1, 49-62.
- Álvarez, Rosario (2020): " O Tesouro do Léxico Patrimonial: proposta metodolóxica para o acervo das variantes xeolectais", en Renato Rodrigues-Pereira & Daniela de Souza Silva Costa (eds.), *Estudos em lexicografia. Aspectos teóricos e prácticos*. Campinas, SP: Pontes Editores, 33-56.
- Álvarez, Rosario & Ernesto X. González Seoane (2020): "Aspectos metodolóxicos da elaboración do corpus Gondomar", en Rosario Álvarez & Ernesto González Seoane (eds.), *Calen barbas, falen cartas. A escrita en galego na Idade Moderna*, Santiago de Compostela: Consello da Cultura Galega, 183-223.
- Fernández Rei, Francisco (1991): "O Instituto da Lingua Galega (1971-1990). Contribución á investigación e á normalización do galego", en Mercedes Brea & Francisco Fernández Rei (coords.), *Homenaxe ó profesor Constantino García*. Santiago de Compostela: Universidade de Santiago de Compostela, vol. 1, 15-46.
- Monteagudo Romero, Henrique (1996): "No vintecinco aniversario do Instituto da Lingua Galega: entrevista con Antón Santamarina", *Grial: revista galega de cultura* 129, 25-42.
- Santamarina, Antón (2011): "O Instituto da Lingua Galega hai 40 anos", *A Trabe de ouro: publicación galega de pensamento crítico* 86, 145-156.
- Sousa, Xulio, Soraya Suárez, Rosa Crujeiras & Laura Calaza. 2020. A GIS-based application for documenting and analysing perceptions about language variation, *Dialectologia* 24, 177-203.



INSTITUTO DA LINGUA GALEGA (ILG)

Moitas grazas!

maria.alvarez.delagranja@usc.gal

elisa.fernandez@usc.gal

