# *Corpus orais: CORILGA*

Xosé Luís Regueira & Elisa Fernández Rei

Instituto da Lingua Galega, Universidade de Santiago de Compostela

# Contidos

Antecedentes e obxectivos

Contidos e estrutura

Ferramentas incorporadas e desenvolvidas

Usos e potencialidades

# Casos prácticos

Exemplos de explotación: cambio linguístico

Exemplos de explotación: variación lingüística

# Antecedentes de CORILGA

Recolla dialectal do ILG desde 1971

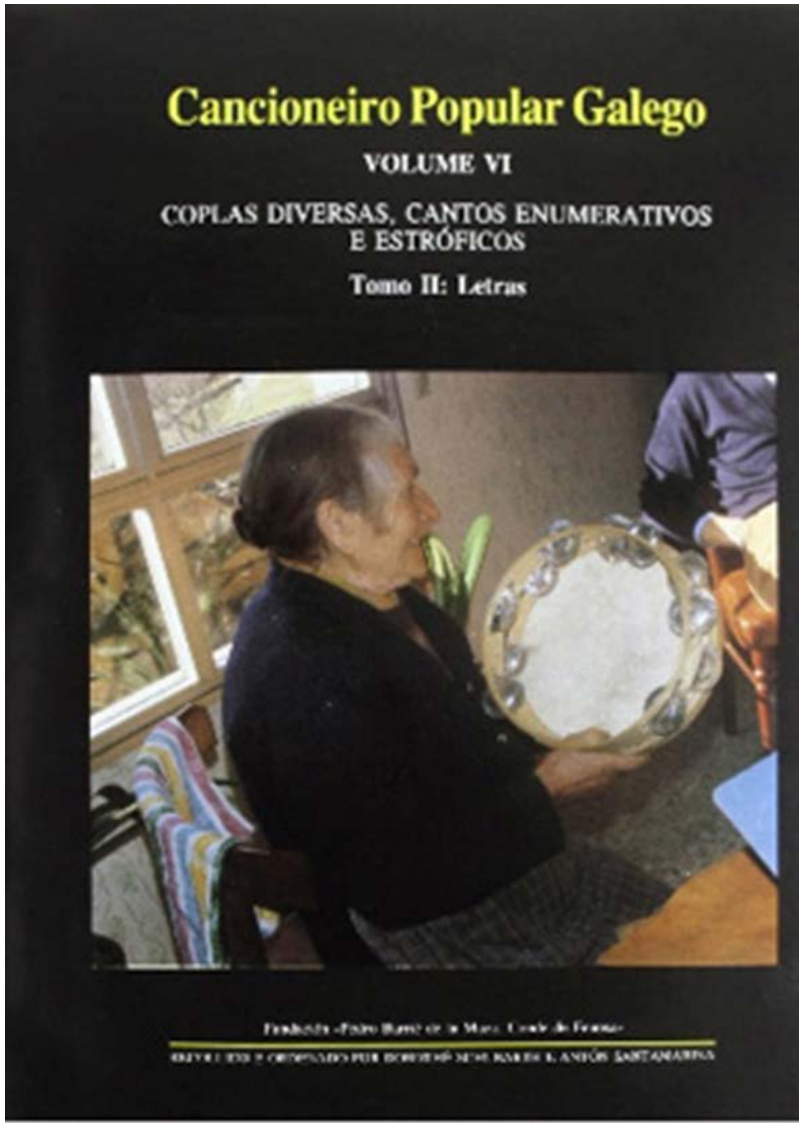
Material dixitalizado

Transcricións ortográficas e fonéticas

Arquivo do Galego Oral (http://ilg.usc.gal/ago/)

# Materiais

- Textos orais representativos de todas as variedades dialectais (maioritariamente falantes NORM)
- Discurso libre, entrevistas semidirixidas
- Narracións, descricións
- Cancións



# Carencias dos corpora e das compilacións orais existentes

- Poucas gravacións de contextos urbanos, de xente nova
- Moi pouca conversa
- Poucas gravacións de boa calidade para análise fonética
- Poucos textos publicados ou dispoñibles
- Sen posibilidade de realizar buscas por non estaren agrupados
- Textos transcritos non aliñados
- Transcricións non anotadas (non lematizadas, sen análise morfosintáctica…)

# Obxectivos

- Dispor dun corpus de gravacións transcritas que permitan estudar a oralidade da lingua galega contemporánea
- Recoller a variación diafásica, diastrática, diacrónica e diatópica
- Facilitar a análise do cambio lingüístico en tempo real e en tempo aparente
- Favorecer a elaboración de estudos humanísticos interdisciplinares, nomeadamente no ámbito lingüístico (análise do discurso, pragmática, morfoloxía, fonética, dialectoloxía etc.).
- Colaborar coa creación e desenvolvemento de tecnoloxías da fala

# Material incorporado a CORILGA

- Gustav Henningsen (diferentes lugares de Galicia, 1965-1967: 100 h sen transcribir)
- Francisco Dubert (Santiago de Compostela, 1994-1996: 20 h de transcrición fonética)
- Xosé Luís Regueira (Vilalba, 1983-1984: 16 h de transcrición fonética)
- Arquivo do Galego Oral (anos 1980-2000, +100 h, transcritas 7 h)
- Manuel Rico (entrevistas RNE-Galicia, anos 1980)
- Noemi Basanta (conversa, 2014)
- Eduardo Louredo (Leiro, 2014)
- Xabier Iglesias (música e conversa, anos 1990)
- Gravacións propias CORILGA (2012 ata a actualidade)
- ....

# Algúns datos de CORILGA

- Horas de gravación: 110:10:00 (+40 prox.)
- Horas transcritas: 57:22:52 (+15 prox.)
- Período: desde os anos 60 ata a actualidade
- Variedades estándar e non estándar
- Falantes de distintas xeracións e diversos niveis socio-culturais

# Tipos de texto

- Oralidade informal (28 h / 34 h)
  - Conversa
  - Entrevista dirixida
  - Monólogo
  - Lectura
  - Literatura oral
- Oralidade formal (18 h / 53 h)
  - Discurso oral
  - Discurso lido
  - Lectura literaria
  - Texto dramático (teatro)

# Tipos de texto

- Medios de comunicación (9 h / 18 h)
  - Informativo
  - Magazin
  - Entrevista
  - Debate
  - Conversa
  - Series
  - Cinema
  - Dobraxe
  - Redes sociais (youtubers)

# Estrutura

- A) Base de datos (MySQL): información sobre a gravación (tipo de texto, tema, data...) e información sociolingüística dos informantes (sexo, idade, residencia etc).
- B) Gravación en formato .wav
- C) Arquivo .eaf (Elan Annotation Format) (Brugman & Russel 2004), con diferentes liñas de anotación para cada informante, aliñadas co audio

# CORILGA

Id: 1

Nome arquivo eaf: OIMO-SURB-AGO-MUROS-01-1996

## 1. DATOS DA GRAVACIÓN

Tipo de texto: 1.5 - Literatura oral

Temas: 4 - Festas e romarías

Hábitat: Semiurbano

Lugar: Muros

Parroquia: Muros

Concello: Muros

Provincia: A Coruña

Contexto:

Córpora oral de procedencia: AGO

Notas:

Data: / / 1996

Responsábel da gravación:

## 2. DATOS SOBRE FALANTES

| Código falante | Sexo | Tramo de id | Idade | Nivel de estu | Profesión | Lugar de nac | Parroquia d | Concello de nacemento | Lugar de res | Parroquia d | Concello de |
|---|---|---|---|---|---|---|---|---|---|---|---|
| H-TI3-GAL-01 | Home | TI3 | 29 | | Mariñeiro | | | | | | |
| | | | | | | | | | | | |

## 3. DATOS SOBRE A TRANSCRICIÓN DA GRAVACIÓN

| | |
|---|---|
| Minutos transcritos | 17 |
| Segundos transcritos | 3 |

| | |
|---|---|
| Minutos totais de gravación | 17 |
| Segundos totais de gravación | 3 |

Transcrición completa ☑

Responsábel da transcrición: Noemi Basanta Llanes

Transcrición revisada ☐

Responsábel da revisión:

## 4. DATOS TOTAIS DE TRANSCRICIÓN DO PROXECTO CORILGA

Horas totais do proxecto transcritas

| | |
|---|---|
| hh_mm_ss | 100:45:11 |

Horas transcritas por corpus

| hh:mm:ss | Corpus |
|---|---|
| 07:11:47 | AGO |
| 00:21:06 | ALGA |
| 00:59:06 | AMPER |
| 00:12:28 | CABR |
| 03:12:11 | CBAS |
| 69:11:42 | CHEN |
| 00:15:25 | CLOU |
| 10:21:37 | CORILGA |
| 05:48:01 | CPRO |

# Liñas de anotación

- Liñas de lingua e tema
- Anotacións:
  - Nivel básico: texto ortográfico
    - Normas de transcrición: quendas, interrupcións… (adapt. de Payrató, 2003)
    - Unidade de transcrición
  - Nivel fonético (AFI)
  - Nivel léxico (palabras e lemas)
  - Nivel morfosintáctico

| | |
|---|---|
| 3.2. Separación de palabras | espazo en branco |
| 3.3. Alongamento dun son | : :: ::: seguido dun espazo<br>Exemplo: texto afectado::: |
| 3.4. Corte abrupto no medio dunha palabra | # sen espazo precedente.<br>Exemplo: texto afect# |
| 3.5. Entoación interrogativa | ¿texto afectado? |
| 3.6. Entoación exclamativa | ¡texto afectado! |
| 3.7. Fin de unidade prosódica | \| precedida e seguida dun espazo.<br>Exemplo: texto \| afectado \| |
| 3.8. Pausa breve ou mediana | \|\| precedida e seguida dun espazo.<br>Exemplo: texto \|\| afectado \|\| |
| 3.9. Pausa de longa duración | \| <segundos> precedida e seguida dun espazo.<br>Exemplo: texto \| <0.5> afectado |
| 3.10. Énfase | Maiúsculas |
| | Exemplo: TEXTO AFECTADO |
| 3.11. Intensidade | |
| Intensidade forte "forte" | {(F) texto afectado} |
| Intensidade moi forte | {(FF) texto afectado} |

# Exemplo de transcrición

H-TI3-GAL-01-ORT bon | compañeiras | compañeiros | delegazóns convidadas |

H-TI3-GAL-01-ORT <0.6> (INH) benvinda:s | a todas e a todos || <0.5> (INH)

H-TI3-GAL-01-ORT a este ato de encerramento | da sétima asembleia | nacional | de nós unidade popular ||

H-TI3-GAL-01-ORT <0.6>

H-TI3-GAL-01-ORT decorreu | ao longo do día de: | de hoxe | <0.6> (INH) con produtivos debates | <0.4> e conclusóns clarificatorias ||

H-TI3-GAL-01-ORT <0.9> (INH)

H-TI3-GAL-01-ORT {<pausa sonora> e::} | non quixese: | deixar pasar esta oportunidade | sin facer menzón | (INH) {<pausa sonora> e:} | a un tema | <0.6> {<pausa sonora> e:::} | arrepiante | a un tema importante | <1.1> {<pausa sonora> e} | que está | que está a contecer | non? ||

H-TI3-GAL-01-ORT <0.4> (INH)

H-TI3-GAL-01-ORT de todas | e de todos | é coñecida | (INH) a nova reforma da lei de seguranza cidadá ||

H-TI3-GAL-01-ORT <0.7> (INH)

H-TI3-GAL-01-ORT agora resulta | <0.4> que TAMBÉN | será delito | ofender | a españa ||


# Ferramentas incorporadas

**Primeiro Corpus: Oralidade formal**

30 arquivos de duración media de 3:50 minutos

Duración Total de 115 minutos (aprox. 2 horas)

**Segundo Corpus: Programas de noticias (TVG)**

10 arquivos de duración media de 34 minutos

Duración Total de 340 minutos (5 horas e 40 min.)

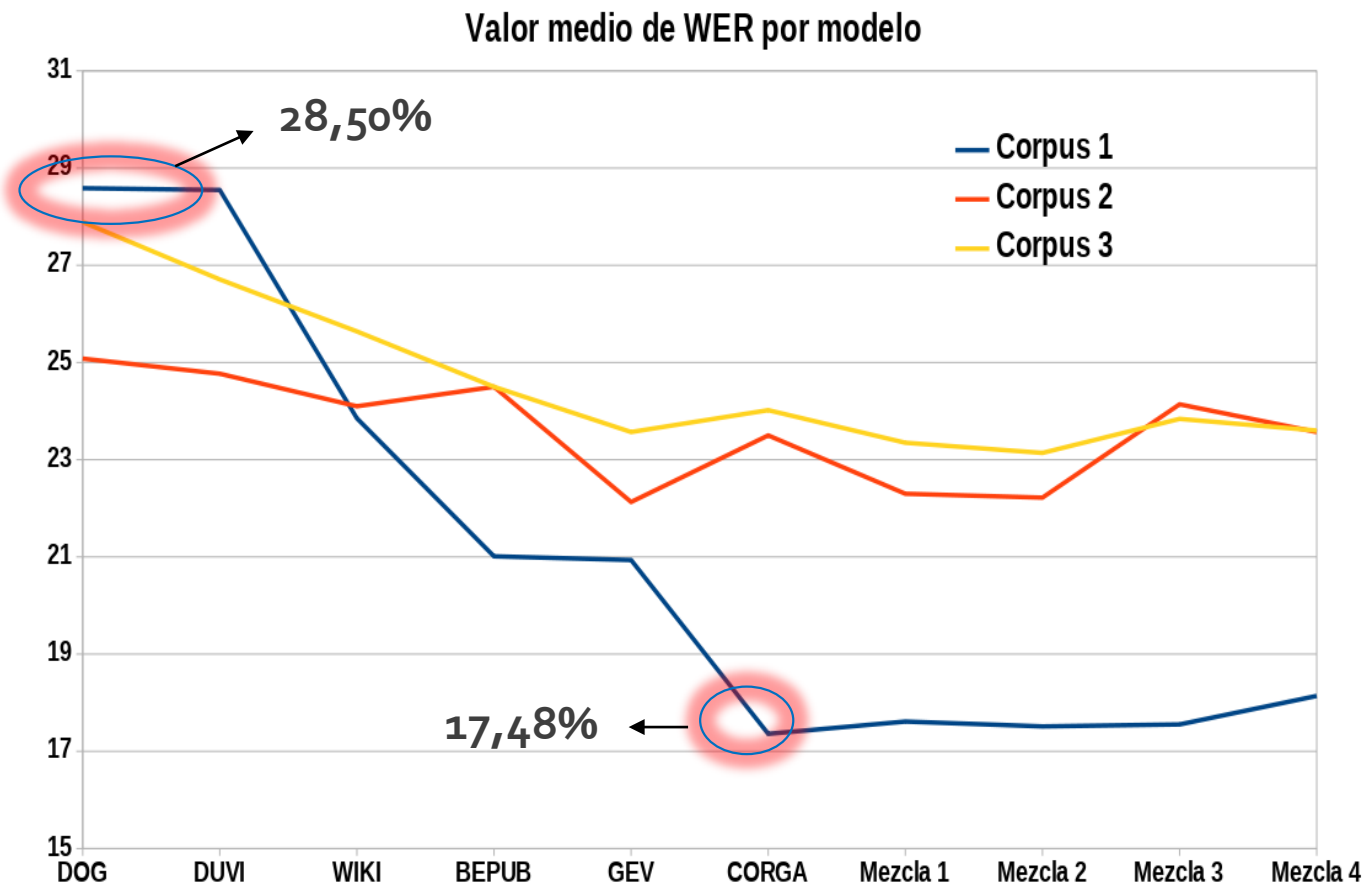
**Terceiro Corpus: TED Talks**

10 arquivos de duración media de 16 minutos

Duración Total de 163 minutos (2 horas e 43 min.)

Adestramento do modelo de linguaxe (Piñeiro et al. 2018): ASR Kaldi , SRI Language Modeling Toolkit

# Avaliación das melloras




Piñeiro et al (2018)

# Posibilidades de explotación do corpus

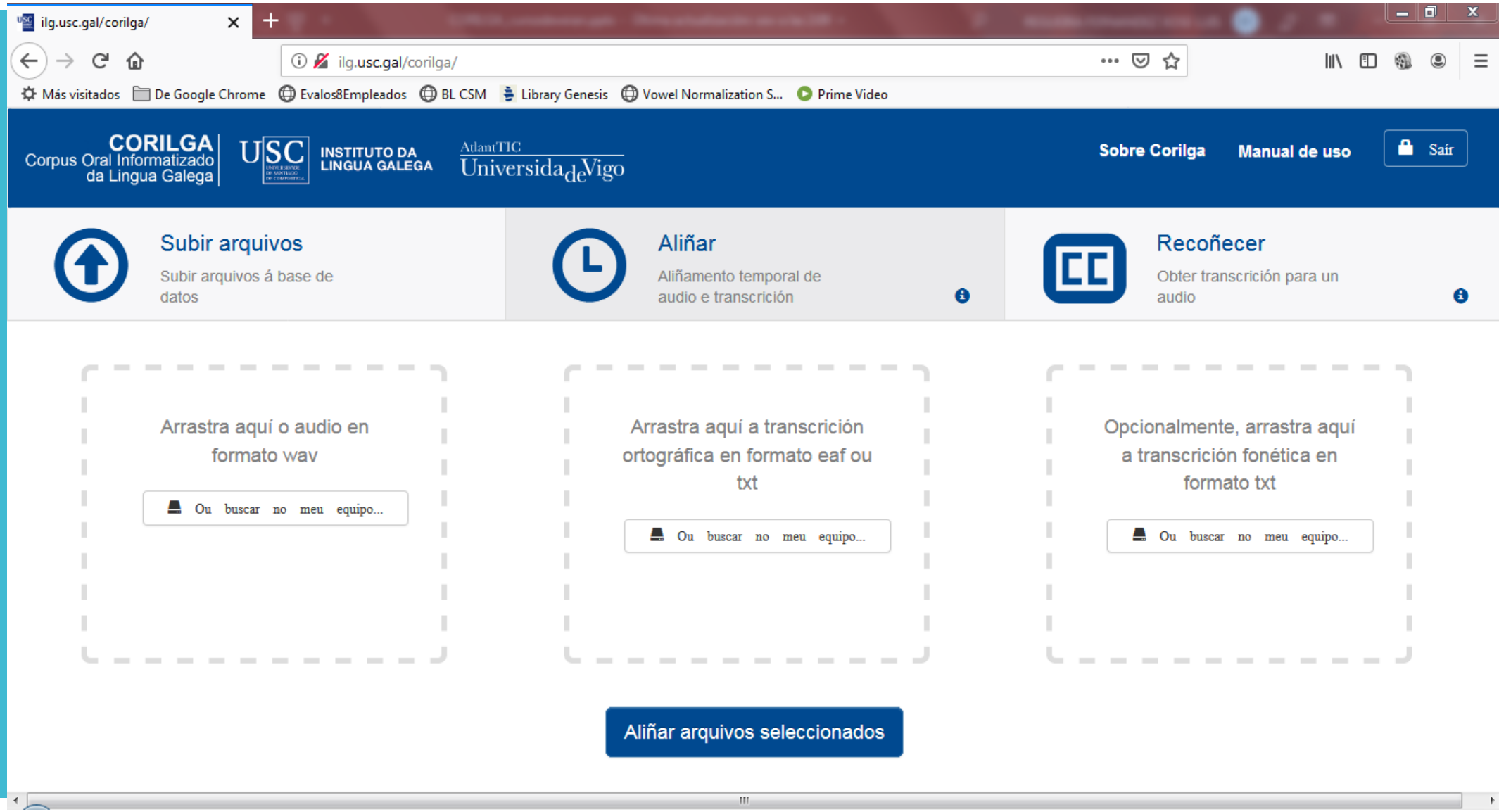
# Debilidades e potencialidades

- **Contra**: fragmentos de fala (limitacións para análise do discurso e da conversa)
  - Necesidade de gran cantidade de tempo en labores de transcrición e corrección (reducida significativamente por medio das ferramentas desenvolvidas). Dificultades de financiamento.
- **Pro**: potente ferramenta para o estudo da variación sociolingüística na lingua actual, así como do cambio nas últimas décadas
  - Contribución ao desenvolvemento e mellora das tecnoloxías da fala
  - Posibilidade de incorporar o español de Galicia (proxecto PRESEGAL) e portugués (Corp-Oral)

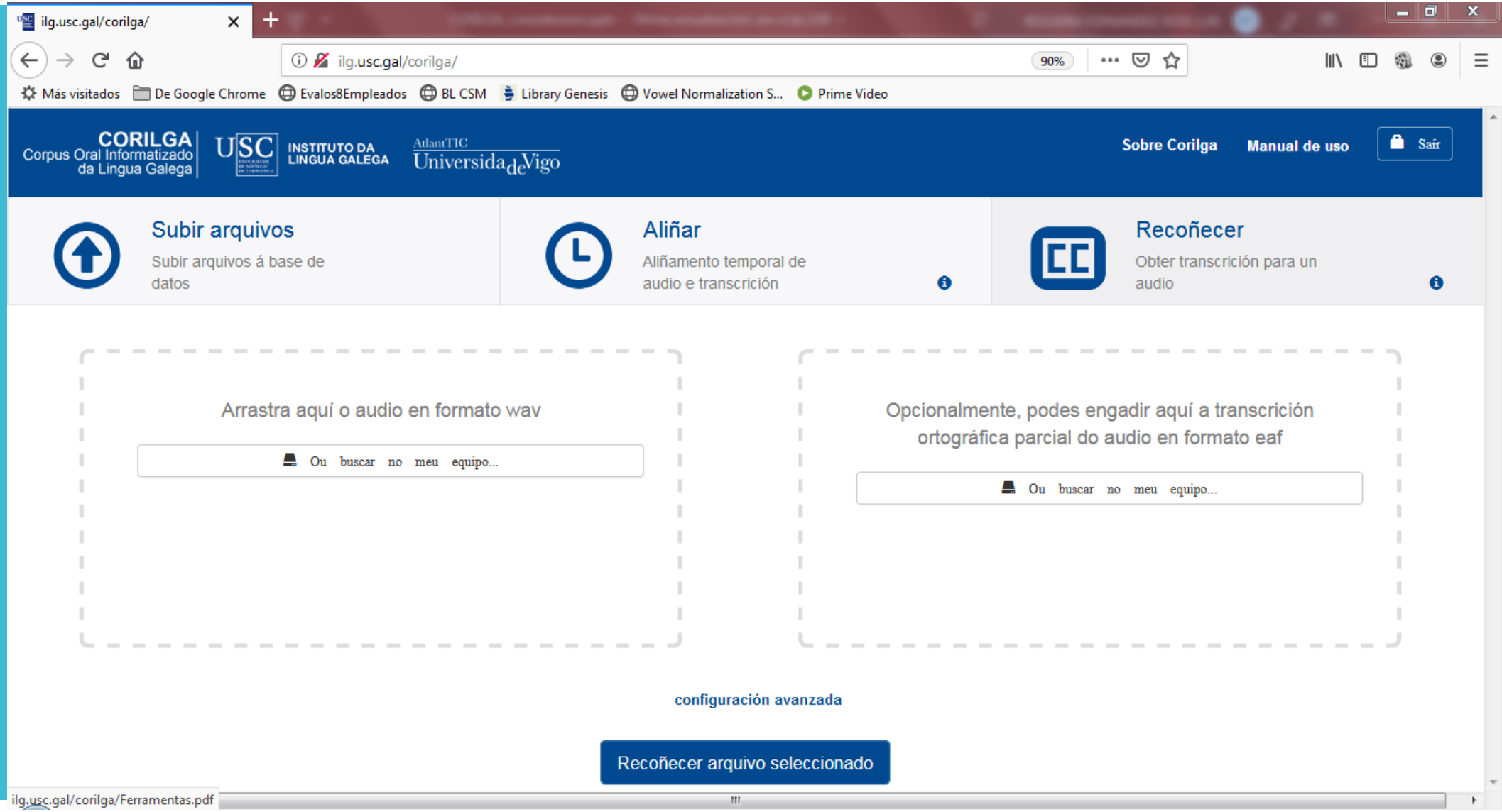
# Funcionamento

- http://ilg.usc.gal/corilga/
- Casos de estudo

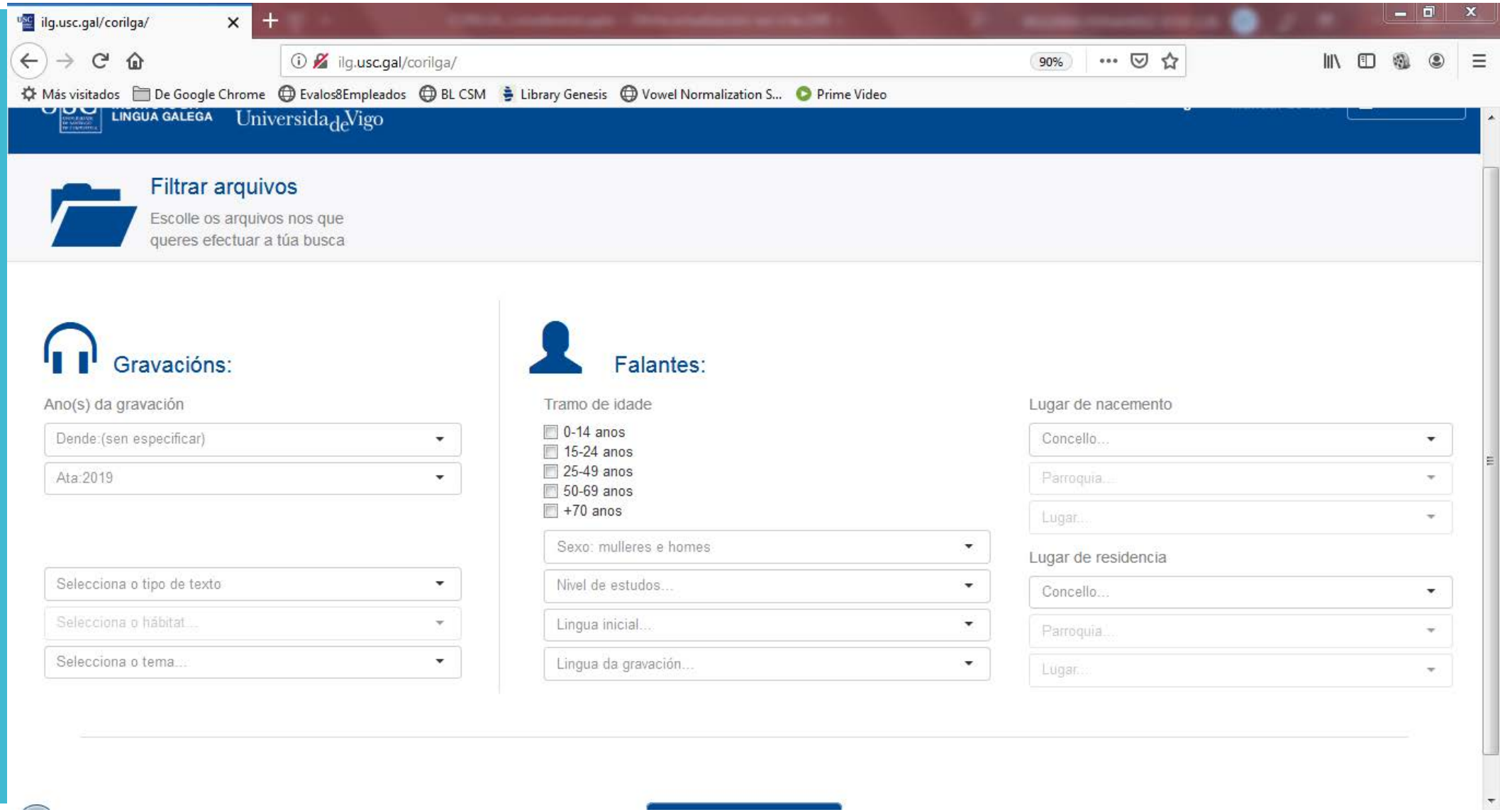
# Páxina de administración: aliñamento



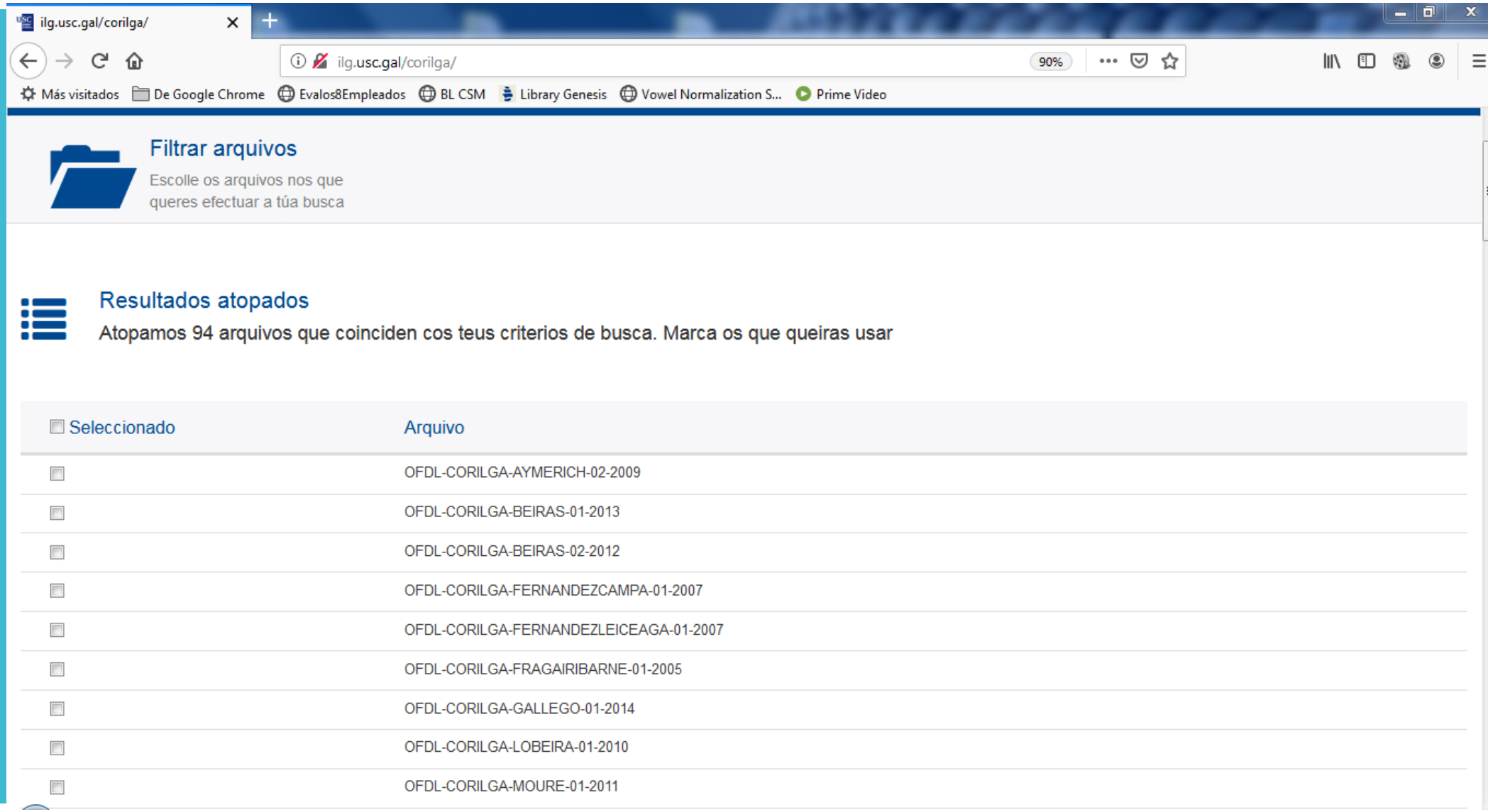
# Páxina de administración: recoñecemento



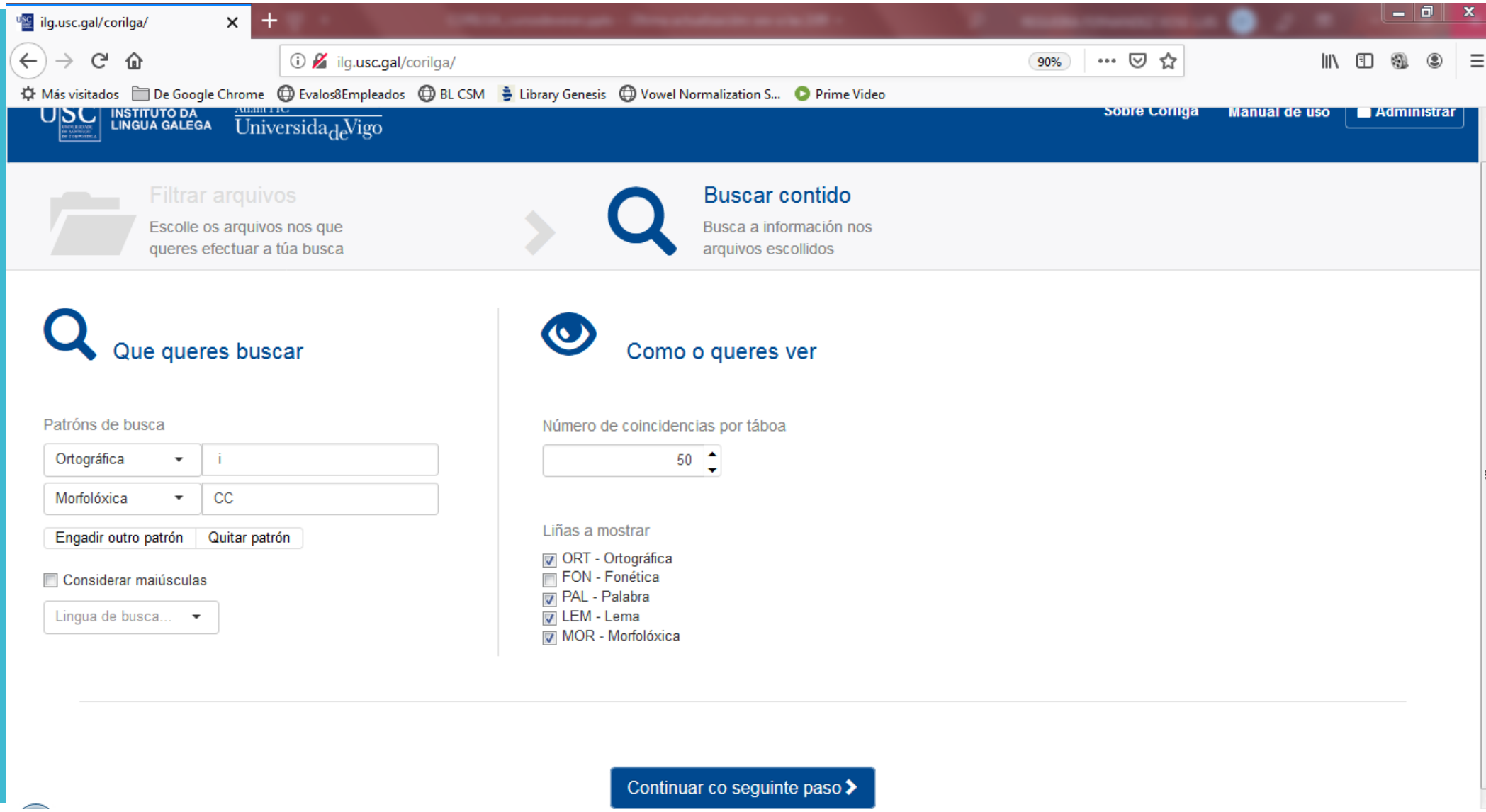
# Páxina de busca: selección de gravacións e de falantes



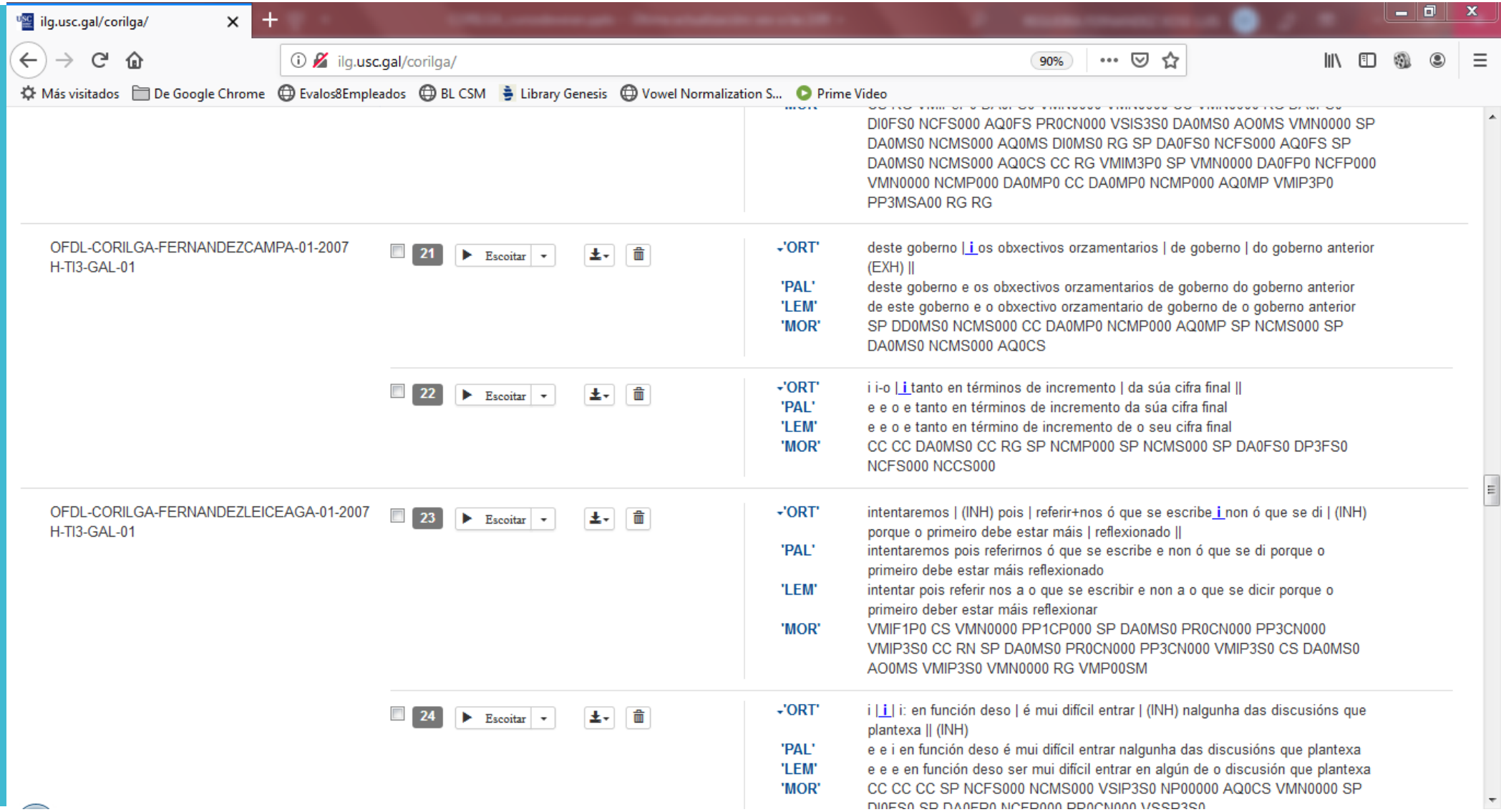
# Páxina de busca: selección de arquivos resultantes da busca anterior



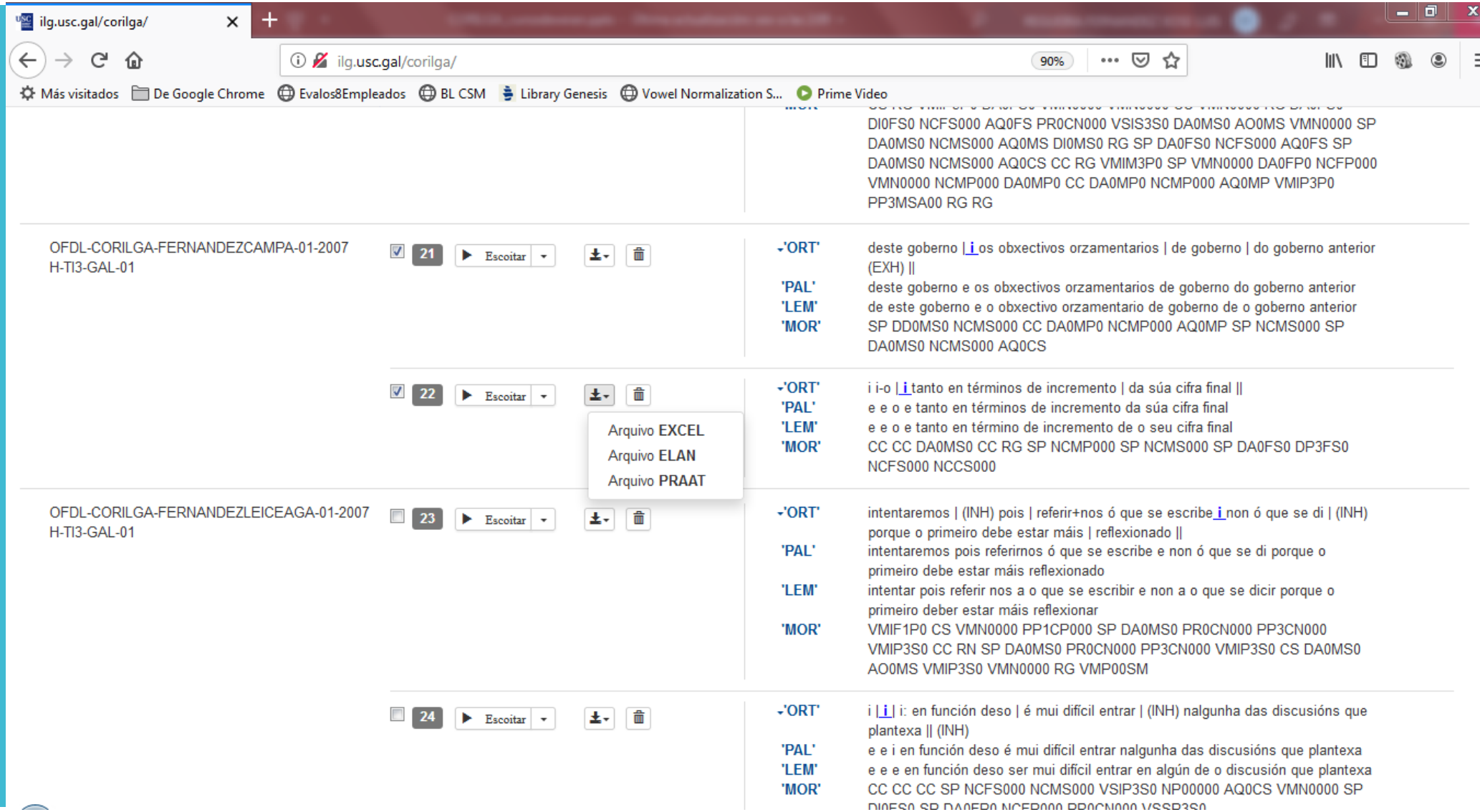
# Páxina de busca: selección de cadea lingüística e outros criterios (posibilidade de buscas combinadas)



# Páxina de busca: resultados da busca (audio e niveis de anotación seleccionados)



# Páxina de busca: posibilidade de baixar resultados en formato excel, elan ou praat



# Referencias


- Brugman, H. & A, Russel (2004). Annotating Multimedia / Multi-modal resources with ELAN. In: *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.*
- Padró, Lluís (2011): Analizadores Multilingües en FreeLing. *Linguamatica*, 3, 2, 13-20.
- Payrató, Lluís (2003): Pragmática, *discurs i llengua oral. Introducció a l'anàlisi funcional de textos.* Barcelona: UOC.
- Piñeiro Martín, A., García-Mateo, C., Docío-Fernández, L., Regueira, X.L. (2018): Estudio sobre el impacto del corpus de entrenamiento del modelo de lenguaje en las prestaciones de un reconocedor de habla. *Procesamiento de Lenguaje Natural* 61, 75-82.
- Povey, D., A. Ghoshal, G. Boulianne , L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer e K. Veselý (2011): "The Kaldi Speech Recognition Toolkit", in *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*. Big Island, Hawaii, 2011 https://www.danielpovey.com/files/2011_asru_kaldi.pdf (consulta 18.05.2019).