

# *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

Exemplos de explotación:  
cambio lingüístico

Exemplos de explotación:  
variación lingüística

Casos prácticos

## Antecedentes de CORILGA

---

Recolla dialectal do ILG desde 1971

---

Material dixitalizado

---

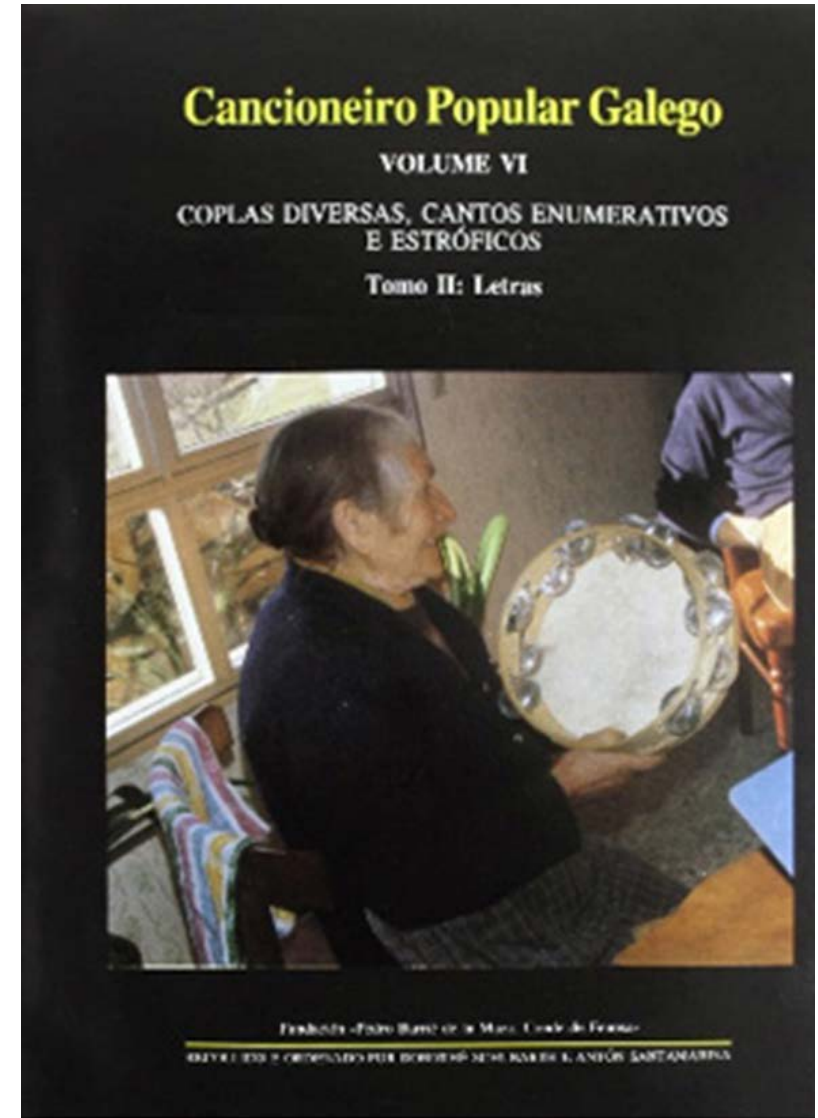
Transcricións ortográficas e fonéticas

---

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 transcripción fonética)
- Xosé Luís Regueira (Vilalba, 1983-1984: 16 h de transcripción 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 dirigida
  - 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) Archivo .eaf (Elan Annotation Format) (Brugman & Russel 2004), con diferentes liñas de anotación para cada informante, aliñadas co audio

Id

Nome arquivo eaf

## 1. DATOS DA GRAVACIÓN

Tipo de texto

Hábitat

Lugar

Parroquia

Concello

Provincia

Córpore oral de procedencia

Data  /  /

Responsábel da gravación

Temas

Contexto

Notas



### 3. DATOS SOBRE A TRANSCRIPCIÓN DA GRAVACIÓN

Minutos transcritos

Segundos transcritos

Minutos totais de gravación

Segundos totais de gravación

Transcripción completa

Responsábel da transcripción

Transcripción revisada

Responsábel da revisión

### 4. DATOS TOTAIS DE TRANSCRIPCIÓN DO PROXECTO CORILGA

Horas totais do proxecto transcritas

hh\_mm\_ss

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    |
| 00:05:50 | CPSO    |

## Liñas de anotación

- Liñas de lingua e tema
- Anotacións:
  - Nivel básico: texto ortográfico
    - Normas de transcripción: quendas, interrupcións... (adapt. de Payrató, 2003)
    - Unidade de transcripción
  - 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 <u>afect#</u>                      |
| 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) <u>texto</u> afectado} |

## Exemplo de transcripción

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 conclusións clarificadoras ||

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 TAMBIÉN | será delito | ofender | a  
españa ||

Aliñador  
texto-  
audio

Lematizador

Recoñecemento  
automático da  
fala (ASR): Kaldi  
(Povey 2011)

Analizador  
morfosint.  
autom.  
(Freeling  
galego)  
(Padró 2010;  
Guinovart &  
Solla 2017)

Transcritor  
fonético  
autom.  
(AFI)

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.)

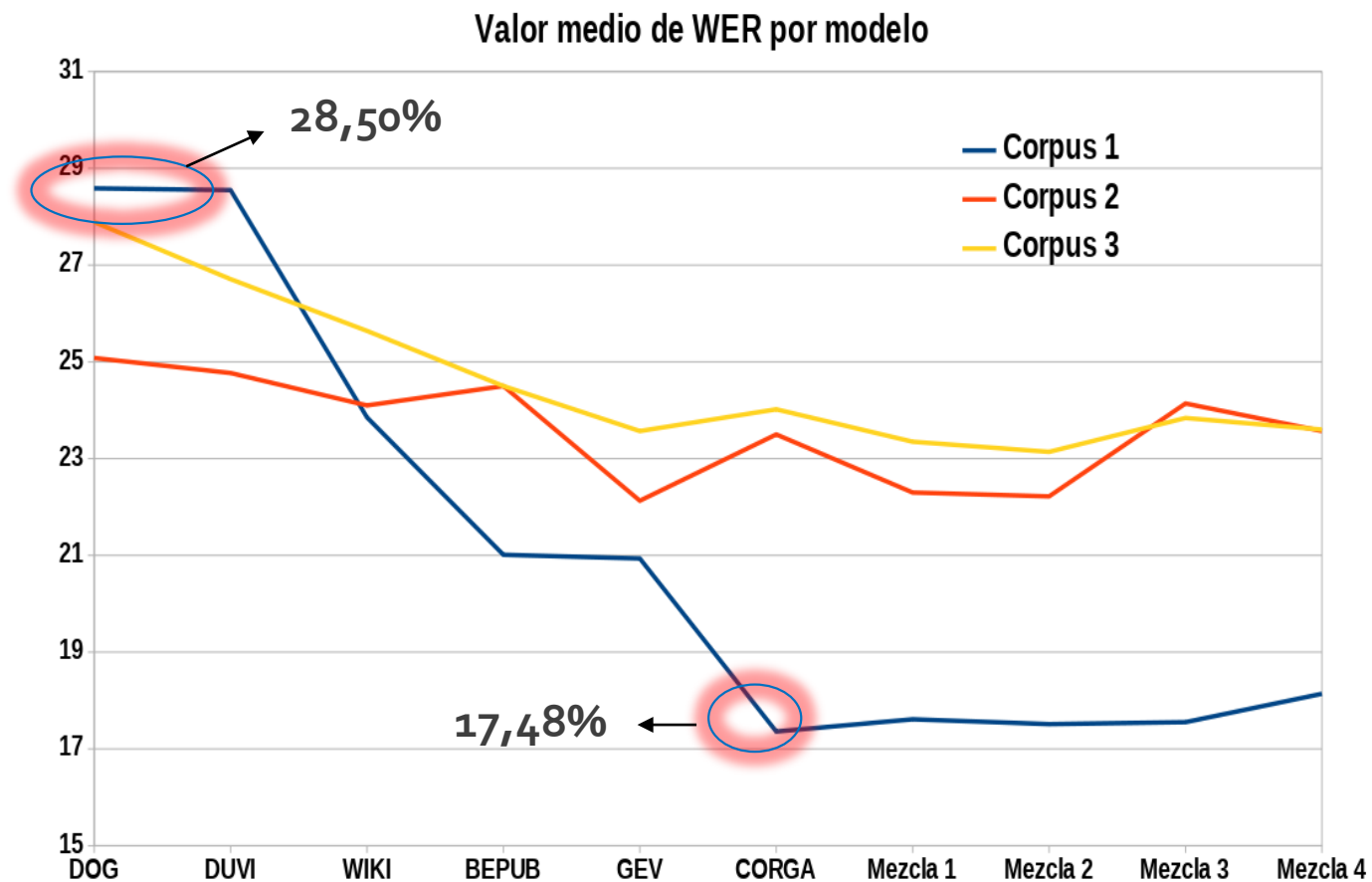


## Terceiro Corpus: TED Talks

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

Cambio en tempo real: mesmo tramo de idade en tempos diferentes

Cambio en tempo aparente: dous tramos de idade no mesmo espazo temporal

Seguimento do mesmo grupo de idade (lonxitudinal)

Variación: rexistro (formal / informal), medios, oral / escrito (lido)...

# Debilidades e potencialidades

- **Contra:** fragmentos de fala (limitacións para análise do discurso e da conversa)
  - Necesidade de gran cantidade de tempo en labores de transcripción 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

The screenshot shows a web browser window with the URL `ilg.usc.gal/corilga/`. The page header includes the CORILGA logo (Corpus Oral Informatizado da Lingua Galega), the USC logo (Universidade de Santiago de Compostela), and the logo for the Instituto da Lingua Galega and AtlantTIC. Navigation links for "Sobre Corilga", "Manual de uso", and "Sair" are visible. The main content area is divided into three columns:

- Subir arquivos**: "Subir arquivos á base de datos". Below this is a dashed box containing the text "Arrastra aquí o audio en formato wav" and a button "Ou buscar no meu equipo...".
- Aliñar**: "Aliñamento temporal de audio e transcripción". Below this is a dashed box containing the text "Arrastra aquí a transcripción ortográfica en formato eaf ou txt" and a button "Ou buscar no meu equipo...".
- Recoñecer**: "Obter transcripción para un audio". Below this is a dashed box containing the text "Opcionalmente, arrastra aquí a transcripción fonética en formato txt" and a button "Ou buscar no meu equipo...".

At the bottom center, there is a large blue button labeled "Aliñar arquivos seleccionados".

# Páxina de administración: recoñecemento

The screenshot shows a web browser window with the URL `ilg.usc.gal/corilga/`. The page features a dark blue header with the CORILGA logo (Corpus Oral Informatizado da Lingua Galega), the USC logo (Universidade de Vigo), and navigation links for 'Sobre Corilga', 'Manual de uso', and 'Saír'. Below the header, there are three main functional buttons: 'Subir arquivos' (Upload files), 'Aliñar' (Align), and 'Recoñecer' (Recognize). The 'Recoñecer' button is highlighted. The main content area contains two large dashed boxes for file uploads. The left box is for audio files in WAV format, with the text 'Arrastra aquí o audio en formato wav' and a search input field 'Ou buscar no meu equipo...'. The right box is for optional orthographic transcription in EAF format, with the text 'Opcionalmente, podes engadir aquí a transcripción ortográfica parcial do audio en formato eaf' and a search input field 'Ou buscar no meu equipo...'. At the bottom, there is a 'configuración avanzada' link and a large blue button labeled 'Recoñecer arquivo seleccionado'. The browser's address bar shows the file path `ilg.usc.gal/corilga/Ferramentas.pdf`.

Páxina de busca:  
selección de  
gravacións e de  
falantes

The screenshot shows a web browser window with the URL `ilg.usc.gal/corilga/`. The page header includes the USC logo and the text "LINGUA GALEGA Universidade de Vigo". Below the header is a section titled "Filtrar arquivos" with the instruction "Escolle os arquivos nos que queres efectuar a túa busca".

The main content area is divided into two columns:

- Gravacións:** Includes a headphones icon and filters for "Ano(s) da gravación" (with "Desde:(sen especificar)" and "Ata:2019" dropdowns), "Selecciona o tipo de texto", "Selecciona o hábitat...", and "Selecciona o tema...".
- Falantes:** Includes a person icon and filters for "Tramo de idade" (with checkboxes for "0-14 anos", "15-24 anos", "25-49 anos", "50-69 anos", and "+70 anos"), "Sexo: mulleres e homes", "Nivel de estudos...", "Lingua inicial...", and "Lingua da gravación...".

On the right side, there are three sections of location filters, each with three dropdown menus:

- Lugar de nacemento:** "Concello...", "Parroquia...", "Lugar..."
- Lugar de residencia:** "Concello...", "Parroquia...", "Lugar..."

Páxina de busca:  
selección de  
arquivos  
resultantes da  
busca anterior

ilg.usc.gal/corilga/

Filtrar arquivos  
Escolle os arquivos nos que queres efectuar a túa busca

Resultados atopados  
Atopamos 94 arquivos que coinciden cos teus criterios de busca. Marca os que queiras usar

| <input type="checkbox"/> Seleccionado | Arquivo                                |
|---------------------------------------|--|
| <input type="checkbox"/>              | OFDL-CORILGA-AYMERICH-02-2009          |
| <input type="checkbox"/>              | OFDL-CORILGA-BEIRAS-01-2013            |
| <input type="checkbox"/>              | OFDL-CORILGA-BEIRAS-02-2012            |
| <input type="checkbox"/>              | OFDL-CORILGA-FERNANDEZCAMPA-01-2007    |
| <input type="checkbox"/>              | OFDL-CORILGA-FERNANDEZLEICEAGA-01-2007 |
| <input type="checkbox"/>              | OFDL-CORILGA-FRAGAIRIBARNE-01-2005     |
| <input type="checkbox"/>              | OFDL-CORILGA-GALLEGO-01-2014           |
| <input type="checkbox"/>              | OFDL-CORILGA-LOBEIRA-01-2010           |
| <input type="checkbox"/>              | OFDL-CORILGA-MOURE-01-2011             |

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

ilg.usc.gal/corilga/

Más visitados De Google Chrome Evalos8Empleados BL CSM Library Genesis Vowel Normalization S... Prime Video

USC INSTITUTO DA LINGUA GALEGA Universidade de Vigo

Sobre Corilga Manual de uso Administrar

Filtrar arquivos  
Escolle os arquivos nos que queres efectuar a túa busca

Buscar contido  
Busca a información nos arquivos escollidos

Que queres buscar

Patróns de busca

Ortográfica i

Morfolóxica CC

Engadir outro patrón Quitar patrón

Considerar maiúsculas

Lingua de busca...

Como o queres ver

Número de coincidencias por táboa

50

Liñas a mostrar

- ORT - Ortográfica
- FON - Fonética
- PAL - Palabra
- LEM - Lema
- MOR - Morfolóxica

Continuar co seguinte paso

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

ilg.usc.gal/corilga/

ilg.usc.gal/corilga/ 90%

Más visitados De Google Chrome Evalos8Empleados BL CSM Library Genesis Vowel Normalization S... Prime Video

DI0FS0 NCFS000 AQ0FS PR0CN000 VSIS3S0 DA0MS0 AO0MS VMN0000 SP  
DA0MS0 NCMS000 AQ0MS DI0MS0 RG SP DA0FS0 NCFS000 AQ0FS SP  
DA0MS0 NCMS000 AQ0CS CC RG VMIM3P0 SP VMN0000 DA0FP0 NCFP000  
VMN0000 NCMP000 DA0MP0 CC DA0MP0 NCMP000 AQ0MP VMIP3P0  
PP3MSA00 RG RG

OFDL-CORILGA-FERNANDEZCAMP-01-2007 H-TI3-GAL-01 21 Escoitar

-'ORT' deste goberno | i os obxectivos orzamentarios | de goberno | do goberno anterior (EXH) ||  
'PAL' deste goberno e os obxectivos orzamentarios de goberno do goberno anterior  
'LEM' de este goberno e o obxectivo orzamentario de goberno de o goberno anterior  
'MOR' SP DD0MS0 NCMS000 CC DA0MP0 NCMP000 AQ0MP SP NCMS000 SP DA0MS0 NCMS000 AQ0CS

OFDL-CORILGA-FERNANDEZLEICEAGA-01-2007 H-TI3-GAL-01 22 Escoitar

-'ORT' i i-o | i tanto en termos de incremento | da súa cifra final ||  
'PAL' e e o e tanto en termos de incremento da súa cifra final  
'LEM' e e o e tanto en término de incremento de o seu cifra final  
'MOR' CC CC DA0MS0 CC RG SP NCMP000 SP NCMS000 SP DA0FS0 DP3FS0 NCFS000 NCCS000

OFDL-CORILGA-FERNANDEZLEICEAGA-01-2007 H-TI3-GAL-01 23 Escoitar

-'ORT' intentaremos | (INH) pois | referir+nos ó que se escribe i non ó que se di | (INH) porque o primeiro debe estar máis | reflexionado ||  
'PAL' intentaremos pois referimos ó que se escribe e non ó que se di porque o primeiro debe estar máis reflexionado  
'LEM' intentar pois referir nos a o que se escribir e non a o que se dicir porque o primeiro deber estar máis reflexionar  
'MOR' VMIF1P0 CS VMN0000 PP1CP000 SP DA0MS0 PR0CN000 PP3CN000 VMIP3S0 CC RN SP DA0MS0 PR0CN000 PP3CN000 VMIP3S0 CS DA0MS0 AO0MS VMIP3S0 VMN0000 RG VMP00SM

OFDL-CORILGA-FERNANDEZLEICEAGA-01-2007 H-TI3-GAL-01 24 Escoitar

-'ORT' i | i | i: en función deso | é mui difícil entrar | (INH) nalgunha das discusións que plantexa || (INH)  
'PAL' e e i en función deso é mui difícil entrar nalgunha das discusións que plantexa  
'LEM' e e e en función deso ser mui difícil entrar en algún de o discusión que plantexa  
'MOR' CC CC CC SP NCFS000 NCMS000 VSIP3S0 NP0000 AQ0CS VMN0000 SP DI0FS0 SP DA0FS0 NCFS000 PR0CN000 VSIS3S0

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

ilg.usc.gal/corilga/

Más visitados De Google Chrome Evalos8Empleados BL CSM Library Genesis Vowel Normalization S... Prime Video

|                                     |    |          |   |    |   |
|-------------------------------------|----|----------|---|----|---|
| <input checked="" type="checkbox"/> | 21 | Escoitar | ↓ | 🗑️ | <p>DI0FS0 NCFS000 AQ0FS PR0CN000 VSIS3S0 DA0MS0 AO0MS VMN0000 SP<br/>                     DA0MS0 NCMS000 AQ0MS DI0MS0 RG SP DA0FS0 NCFS000 AQ0FS SP<br/>                     DA0MS0 NCMS000 AQ0CS CC RG VMIM3P0 SP VMN0000 DA0FP0 NCFP000<br/>                     VMN0000 NCMP000 DA0MP0 CC DA0MP0 NCMP000 AQ0MP VMIP3P0<br/>                     PP3MSA00 RG RG</p>   |
| <input checked="" type="checkbox"/> | 22 | Escoitar | ↓ | 🗑️ | <p>-'ORT' deste goberno   <u>i</u> os obxectivos orzamentarios   de goberno   do goberno anterior (EXH)   </p> <p>'PAL' deste goberno e os obxectivos orzamentarios de goberno do goberno anterior</p> <p>'LEM' de este goberno e o obxectivo orzamentario de goberno de o goberno anterior</p> <p>'MOR' SP DD0MS0 NCMS000 CC DA0MP0 NCMP000 AQ0MP SP NCMS000 SP DA0MS0 NCMS000 AQ0CS</p>   |
| <input checked="" type="checkbox"/> | 23 | Escoitar | ↓ | 🗑️ | <p>-'ORT' i i-o   <u>i</u> tanto en termos de incremento   da súa cifra final   </p> <p>'PAL' e e o e tanto en termos de incremento da súa cifra final</p> <p>'LEM' e e o e tanto en término de incremento de o seu cifra final</p> <p>'MOR' CC CC DA0MS0 CC RG SP NCMP000 SP NCMS000 SP DA0FS0 DP3FS0 NCFS000 NCCS000</p>  |
| <input type="checkbox"/>            | 24 | Escoitar | ↓ | 🗑️ | <p>-'ORT' intentaremos   (INH) pois   referir+nos ó que se escribe <u>i</u> non ó que se di   (INH) porque o primeiro debe estar máis   reflexionado   </p> <p>'PAL' intentaremos pois referimos ó que se escribe e non ó que se di porque o primeiro debe estar máis reflexionado</p> <p>'LEM' intentar pois referir nos a o que se escribir e non a o que se dicir porque o primeiro deber estar máis reflexionar</p> <p>'MOR' VMIF1P0 CS VMN0000 PP1CP000 SP DA0MS0 PR0CN000 PP3CN000 VMIP3S0 CC RN SP DA0MS0 PR0CN000 PP3CN000 VMIP3S0 CS DA0MS0 AO0MS VMIP3S0 VMN0000 RG VMP00SM</p> |
| <input type="checkbox"/>            | 24 | Escoitar | ↓ | 🗑️ | <p>-'ORT' i   <u>i</u>   i: en función deso   é mui difícil entrar   (INH) nalgunha das discusións que plantexa    (INH)</p> <p>'PAL' e e i en función deso é mui difícil entrar nalgunha das discusións que plantexa</p> <p>'LEM' e e e en función deso ser mui difícil entrar en algún de o discusión que plantexa</p> <p>'MOR' CC CC CC SP NCFS000 NCMS000 VSIP3S0 NP00000 AQ0CS VMN0000 SP</p>  |

Arquivo EXCEL  
 Arquivo ELAN  
 Arquivo 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](https://www.danielpovey.com/files/2011_asru_kaldi.pdf) (consulta 18.05.2019).