USC INSTITUTO DA LINGUA GALEGA

atlanTTic research center for Telecommunication Technologies

Universidade de Vigo

CORILGA

un corpus anotado multinivel para estudiar la variación y el cambio en la lengua hablada

XOSÉ LUÍS REGUEIRA* & CARMEN GARCÍA-MATEO**

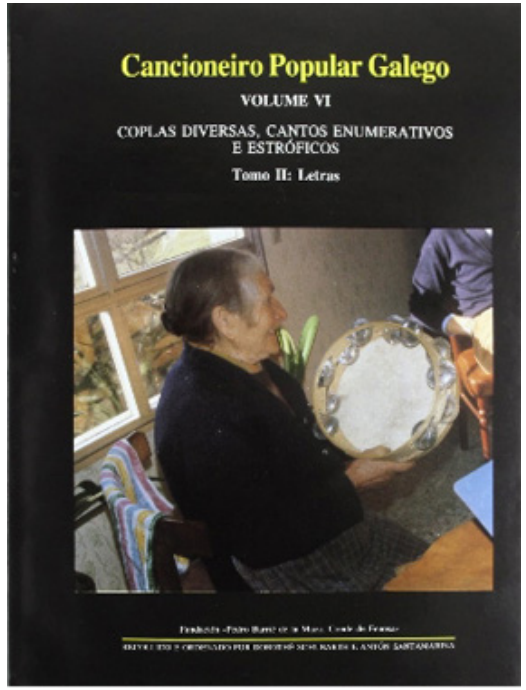
*INSTITUTO DA LINGUA GALEGA – UNIV. DE SANTIAGO DE COMPOSTELA

**ATLANTTIC – UNIV. DE VIGO

LingCor2018 – 1st International Workshop on Spoken Corpus Linguistics
Universitat de València, December 13-14

Antecedentes

- Gran cantidad de textos orales representativos de todas las variedades dialectales (mayoritariamente hablantes NORM)
- Discurso libre, entrevistas semidirigidas
- Narraciones, descripciones
- Canciones
- Elevado número de textos transcritos
- Arquivo do Galego Oral (http://ilg.usc.es/ago/)



Limitaciones

- Pocas grabaciones de contextos urbanos, de gente joven
- Muy poca conversación
- Pocas grabaciones de buena calidad para análisis fonético
- Pocos textos publicados o disponibles
- No permiten hacer búsquedas (textos no agrupados)
- Textos transcritos no alineados
- Transcripciones sin anotación morfosintáctica, no lematizadas

Objetivos

- Un corpus de grabaciones transcritas que permita estudiar la oralidad contemporánea
- Recoger la variación (no solo diatópica)
- Facilitar el análisis del cambio lingüístico en tiempo aparente y en tiempo real
- Favorecer la elaboración de estudios interdisciplinares, esp. lingüísticos (morfología, fonética, dialectología, análisis del discurso, pragmática, etc.).
- Contribuir a la creación y al desarrollo de tecnologías del habla

Algunos datos de CORILGA

- Horas de grabación: 110:10:00
- Horas transcritas: 57:22:52
- Período: desde 1965 hasta la actualidad
- Registro oral formal e informal, variedades estándar y no estándar
- Hablantes de generaciones diferentes e niveles socio-culturales distintos

Córpora

- Gustav Henningsen (diferentes lugares de Galicia, 1965-1967: 100 h sin transcr.)
- Francisco Dubert (Santiago de Compostela, 1994-1996: 20 h transcr. fonética)
- Xosé Luís Regueira (Vilalba, 1983-1984: 16 h transcr. fonética)
- Arquivo do Galego Oral (años 1980-2000, +100 h, transcr. 7 h)
- Manuel Rico (entrevistas RNE-Galicia, años 1980)
- Noemi Basanta (conversación, 2014)
- Eduardo Louredo (Leiro, 2014)
- Xabier Iglesias (música y conversación, años 1990)
- Grabaciones propias CORILGA (2012 hasta la actualidad)

….

Tipos de texto

- Oralidad informal (28 h / 34 h)
  - Conversación
  - Entrevista dirigida
  - Monólogo
  - Lectura
  - Literatura oral
- Oralidad formal (18 h / 53 h)
  - Discurso oral
  - Discurso leído
  - Lectura literaria
  - Texto dramático (teatro)

Tipos de texto

- Medios de comunicación (9 h / 18 h)
  - Informativo
  - Magazin
  - Entrevista
  - Debate
  - Conversación
  - Serie
  - Cinema
  - Doblaje
  - Redes sociales (youtubers)

Datos sociolingüísticos

- Edad
- Nivel de estudios
- Sexo
- Lengua inicial
- Lengua de la grabación
- Lugar de nacimiento
- Lugar de residencia

Estructura

A) Base de datos (MySQL): información sobre la grabación (tipo de texto, lengua, fecha) e información sociolingüística de los informantes (sexo, edad, residencia, etc).

B) Grabación en formato .wav

C) Archivo .eaf (Elan Annotation Format) (Brugman & Russel 2004), con diferentes líneas de anotación para cada informante, alineadas con el audio:

- Ortográfica: transcripción análisis de la conversación (adapt. de Payrató, 2003)
- Fonética (AFI)
- Palabra
- Lema
- Etiqueta morfosintática
- Líneas de lengua (gallego, español) y tema

Posibilidad de búsquedas combinadas (forma / lema + etiqueta, p.e.)

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

Estructura

A) Base de datos (MySQL): información sobre la grabación (tipo de texto, lengua, fecha) e información sociolingüística de los informantes (sexo, edad, residencia, etc).

B) Grabación en formato .wav

C) Archivo .eaf (Elan Annotation Format) (Brugman & Russel 2004), con diferentes líneas de anotación para cada informante, alineadas con el audio:

- Ortográfica: transcripción análisis de la conversación (adapt. de Payrató, 2003)
- Fonética (AFI)
- Palabra
- Lema
- Etiqueta morfosintática
- Líneas de lengua (gallego, español) y tema

Posibilidad de búsquedas combinadas (forma / lema + etiqueta, p.e.)

3.2. Separación de palabras	espazo en branco
3.3. Alongamento dun son	: :: ::: seguido dun espazo Exemplo: texto afectado:::
3.4. Corte abrupto no medio dunha palabra	# sen espazo precedente. 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. Exemplo: texto | afectado |
3.8. Pausa breve ou mediana	|| precedida e seguida dun espazo. Exemplo: texto || afectado ||
3.9. Pausa de longa duración	| <segundos> precedida e seguida dun espazo. 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}

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 ||

Estructura

A) Base de datos (MySQL): información sobre la grabación (tipo de texto, lengua, fecha) e información sociolingüística de los informantes (sexo, edad, residencia, etc).

B) Grabación en formato .wav

C) Archivo .eaf (Elan Annotation Format) (Brugman & Russel 2004), con diferentes líneas de anotación para cada informante, alineadas con el audio:

- Ortográfica: transcripción análisis de la conversación (adapt. de Payrató, 2003)
- Fonética (AFI)
- Palabra
- Lema
- Etiqueta morfosintática
- Líneas de lengua (gallego, español) y tema

Posibilidad de búsquedas combinadas (forma / lema + etiqueta, p.e.)

Posibilidades de explotación del corpus

- Cambio en tiempo real: mismo tramo de edad en tiempos diferentes
- Cambio en tiempo aparente: dos tramos de edad en el mismo espacio temporal
- Seguimiento del mismo grupo de edad (longitudinal)
- Variación: registro (formal / informal), medios, oral / escrito (leído)…

Herramientas desarrolladas / incorporadas

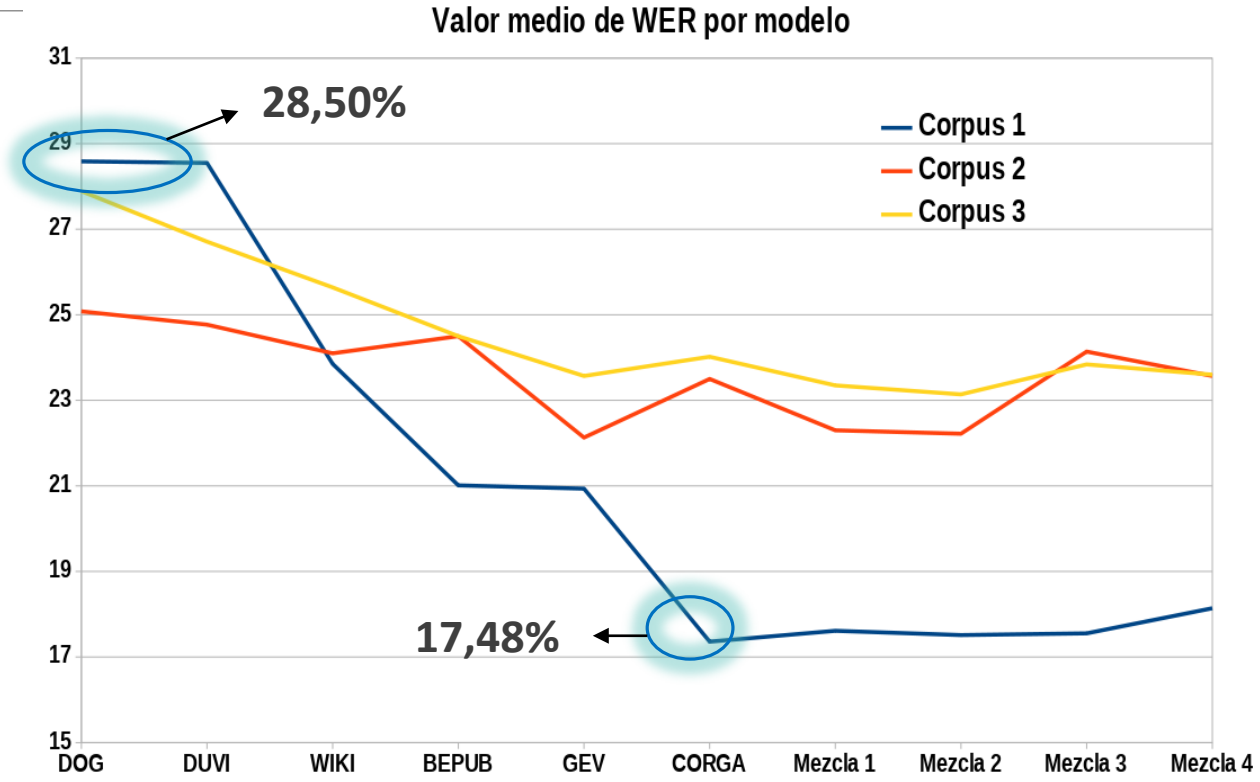
- Alineador texto-audio (palabra, segmento)
- Reconocimiento automático del habla (ASR): Kaldi Speech Recognition Toolkit (Povey, 2011). KALDI
  - Archivo eaf (ELAN), con índice de nivel de confianza
- Transcriptor fonético automático (IPA) (basado en SAMPA)
- Analizador morfosintáctico automático (Freeling galego) (Padró 2010; Guinovart & Solla 2017)
- Lematizador

Desarrollo de tecnologías del habla

- Mejoras del reconocedor de habla: entrenamiento del modelo de lenguaje (Piñeiro et al. 2018): ASR Kaldi , SRI Language Modeling Toolkit

- **Primer Corpus: Oralidad formal**
  - 30 archivos de duración media de 3:50 minutos
  - Duración Total de 115 minutos (aprox. 2 horas)
- **Segundo Corpus: Programas de noticias (TVG)**
  - 10 archivos de duración media de 34 minutos
  - Duración Total de 340 minutos (5 horas y 40 min.)
- **Tercer Corpus: TED Talks**
  - 10 archivos de duración media de 16 minutos
  - Duración Total de 163 minutos (2 horas y 43 min.)

Desarrollo de tecnologías del habla



Piñeiro et al (2018)

Funcionamiento

http://ilg.usc.gal/corilga/

[vídeos aliñamento e recoñecemento]

Buscas: rmos / formal

Debilidades y potencialidades

- **Contra**: fragmentos de habla (limitaciones para análisis del discurso y de la conversación)
  - Demasiado ambicioso? Necesidad de gran cantidad de tiempo en labores de transcripción y corrección (reducida significativamente por medio de las herramientas desarrolladas). Dificultades de financiación.
- **Pro**: potente herramienta para el estudio de la variación sociolingüística en la lengua actual así como del cambio en las últimas décadas
  - Contribución al desarrollo y mejora de las tecnologías del habla
  - Posibilidad de incorporar el español de Galicia (proyecto PRESEGAL) y portugués (Corp-Oral)

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.

Gràcies