O proxecto CORILGA: Corpus Oral Informatizado da Lingua Galega

Xosé Luís Regueira
Elisa Fernández Rei
Francisco Cidrás Escáneo
Xulio Sousa Fernández

Materiais orais do ILG

- Gran cantidade de textos orais representativos de todas as variedades dialectais
- Discurso libre, entrevistas semidirixidas
- Narracións, descricións
- Cancións
- Elevado número de textos transcritos

Corpora e materiais dispoñibles

- Arquivo do Galego Oral (ILG): subcorpora:
  - A) 2000 horas de gravación (desde 1974), en proceso de dixitalización.
    - Fichadas e clasificadas. Parcialmente transcritas.
    - Predominio de áreas rurais, idade media-avanzada.
    - Predominio de entrevista semidirixida (narración, descrición).
    - Interese etnográfico e histórico.
    - Valiosas para el estudio de léxico, gramática, discurso.
    - Baixa calidade (grav. casete dixitalizada)

Corpora e materiais dispoñibles

- Arquivo do Galego Oral (ILG): subcorpora:
  - B) Varias ducias de textos cultos e formais (conferencias…)
    - Boa calidade
    - Non transcritas
  - C) *Cancioneiro Popular Galego* (Schubart & Santamarina 1978-1983): 350 h de cancións e romances.

Corpora e materiais dispoñibles

- Arquivo do Galego Oral (ILG):
  - Publicados:
  - *A nosa fala* (Fernández Rei & Hermida): 1996, 2003
    - http://consellodacultura.org/arquivos/asg/anosafala.php
  - *Arquivo do Galego Oral* (Fernández Rei, dir.):
    - http://ilg.usc.es/ago/

ARQUIVO SONORO DE GALICIA

asonoro@consellodacultura.org
Tel: +34 881 995 147

feito por culturagalega.org

A Nosa Fala
Bloques e Áreas Lingüísticas do Galego (versión interactiva con audio)

Inicio

NAVEGA POLOS BLOQUES LINGUÍSTICOS
Bloque Occidental
Bloque Oriental
Bloque Central

ETNOTEXTOS. BLOQUE OCCIDENTAL
1975 | **Noia** | *labrego, 76*
1975 | **A Estrada** | *labrego, 55*
1975 | **Mondariz** | *labrego, 62*
1976 | **Cambados** | *mariñeiro, 49*
1979 | **Salvaterra de Miño** | *labrega, 40*
1980 | **Cerdedo** | *labrega, 55*
1981 | **O Rosal** | *labrego, 64*
1982 | **Teo** | *labrega, 60*
1983 | **Marín** | *muiñeira, 60*
1985 | **Mazaricos** | *labrega, 46*
1986 | **Tomiño** | *veciña, 66*
1993 | **Coristanco** | *labrega, 85*
1994 | **Camariñas** | *mariñeiro, 63*
1994 | **Cangas do Morrazo** | *estudante, 14*
1995 | **Ribeira** | *mariñeiro, 24*

As diversións e a mocidade de antes

Lugar: **Roo - Santa María de Roo** .
Concello:**Noia**

Observacións:
Informante: Un veciño, 76 anos, labrego.
Data: Xullo de 1975.
Gravación: M. González González.
Transcrición: F. Fernández Rei.
Fonte: AGO-ALGa.

// DESCARGA AQUÍ O ARQUIVO EN MP3

Transcripción//

È poise na, nas casas façían, formábase un baile nunha sa[la], nunha casa è coa misma prinçipiábaçe a toca-la pandèreta è veña baile: a ghòta, a muiñeira, o pasodòble, o valse. È en fin, tódolos bailes que había: dança, maçurca, todo que aghora non había estes bailes modèrnos de, de abraçarçe tanto, de apretarçe tanto. Eço non, çeparados un de outro. È entonçes si, aquelo daba gusto. Ademais, as mullères, daquèla iban coas saiñas asta alá abaixo, a casi a rastro do chan; pèro despois, este, a bailar, poise non nos freghábamos nada, estábamos separados, è, en fin. Non había bicos tampouco no baile. Eso, os bicos èra cando ìbamos, por eghèmplo, ìbamos açí è estábamos nun çitio tal; pèro cando çe lle daba un bico a unha mullèr xa èra unha couça moi sèria: ¡Diòs nos libre deço! Daquèla a

roxecto AGO

tica do galego oral foi unha das principais do Instituto da Lingua Galega (ILG) da go de Compostela desde a súa fundación onsel do traballo dialectal que se iniciara ia Románica compostelá en 1966. Ese fundamentalmente, para redactar s e léxicas de parroquias e concellos ou e fauna de todo o litoral galego, ou ben caso do material do *Atlas Lingüístico* omezou a editarse en 1990.

practicamente ininterrompido desde o lade, diversos investigadores do ILG de etnotextos en moitos puntos do alego, á vez que dirixiron traballos os nos que o material fundamental para stra do galego oral con un ou varios

unto con outras moitas en depósito de stitúen o proxecto *Arquivo do Galego* pretende construir un corpus de interese udar a situación da lingua oral e asemade oñecemento da sociedade galega a través

Concello: Guitiriz
Parroquia: VILARES (SAN VICENTE)
Lugar: As Reixas
Tema: Acontecementos e personaxes históricos
Subtema: A guerra civil e a posguerra
Informante/s: Muller 91 anos - xubilada
Gravación de: Francisca Pilar Miragaya Fernández
Ano gravación: 1995

Semifonolóxica Ambas Estandarizada

0:00 1:48

Audio 30

Transcrición semifonolóxica

Cando vèu a ghèrra, o, os hòmes ibanse pr'â guèrra è as mullères choraban è traballaban, è viñ'à xunta da noite è cèrra-las pòrtiñas que decían qu'habíascapados è, è, è, è cèrrábans'as pòrtas, eso. Tábase cabo do lumiño è todo cèrrado; è despois, viñan moitos pòbres pidire, porque..., dos puèblos, viñan pòbres da Curuña è de, è doutros lados è con pequeniños, è miña mamai en pa-descanse, nosoutros cociamos unha fornada de pan, è miña mamai en pa-descanse quedaba con tres nenas qu'èu tiña è, è cando viñä â noite non habia bocado de pan, que x'hàbia que encetar outro, que llo daba todo a cantos pòbres viña, a cada un o seu curruchiño de pan, cada un o seu curruchiño de pan.
È despois dixo, tiñamos espigas de mainzo; è

Transcrición estandarizada

Cando veu a guerra, os homes ían para a guerra e as mulleres choraban e traballaban, e viña a xunta da noite e cerrar as portiñas que dicían que habia escapados, e cerrábanse as portas, iso. Estábase cabo do lumiño e todo cerrado; e despois, viñan moitos pobres pedir, porque das vilas, viñan pobres da Coruña e doutros lados e con pequeniños, e miña mamá en paz descanse, nosoutros cociamos unha fornada de pan, e miña mamá en paz descanse quedaba con tres nenas que eu tiña e cando viña á noite non habia bocado de pan, que xa habia que encetar outro, que lleo daba todo a cantos pobres viñan, a cada un o seu curruchiño de pan, cada un o seu curruchiño de pan.
E despois dixo, tiñamos espigas de mainzo; e

Outros corpora

- Gustav Henningsen, 1964-1968. 650 gravacións, 104 h (ILG). En fase de transcrición.
- Proxecto AMPER (ILG): 40 gravacións, entrevista semidirixida e Map-Task.
- Proxecto Prosodia do discurso formal: 20 gravacións de conferencias, charlas…
- Outros en institucións e privados. Ex. Arquivo Sonoro de Galicia (http://consellodacultura.org/arquivos/asg/)

Córpora orais: déficit

- Poucas gravacións de contextos urbanos, de xente nova
- Moi pouca conversación
- Poucas gravacións de boa calidade para análise fonética
- Poucos textos publicados
- Non permiten buscas en córpora extensos
- Textos transcritos non aliñados
- Transcricións non anotadas

CORILGA: Corpus Oral Informatizado da Lingua Galega
Obxectivos:

- Corpus oral aberto, extenso e utilizable
- Seleccionar, ampliar e completar:
  - Variedades escasamente representadas: falas urbanas, fala xuvenil
  - Conversación
- Incrementar a funcionalidade:
  - Transcricións aliñadas e anotadas
  - Permitir buscas en córpora extensos
  - Acceso público

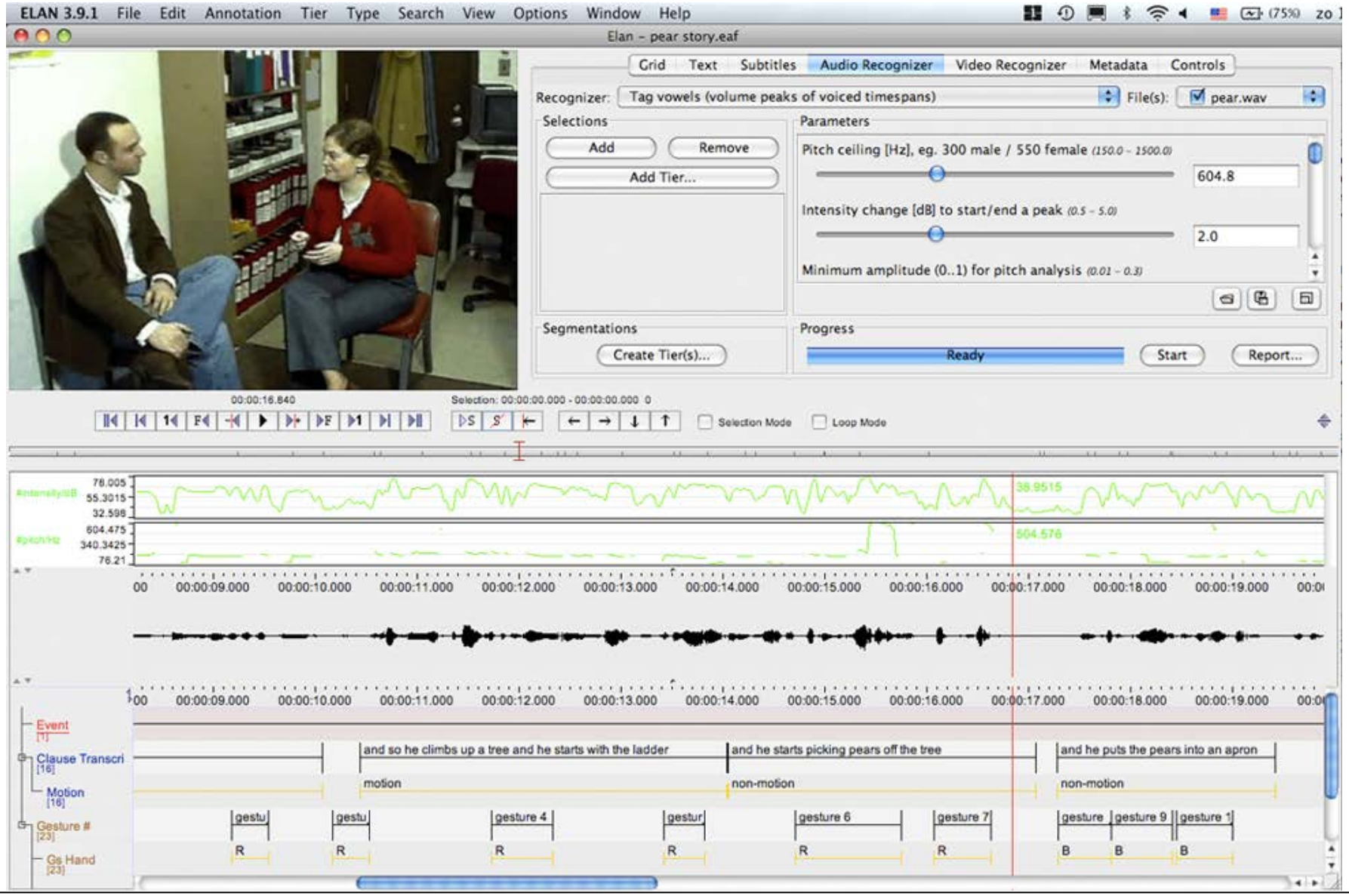
CORILGA: Aplicacións

- Primeiras aplicacións:
  - Prosodia
  - Fonética
  - Variación e cambio lingüístico
- Outras aplicacións:
  - Discurso
  - Sintaxe
  - Pragmática

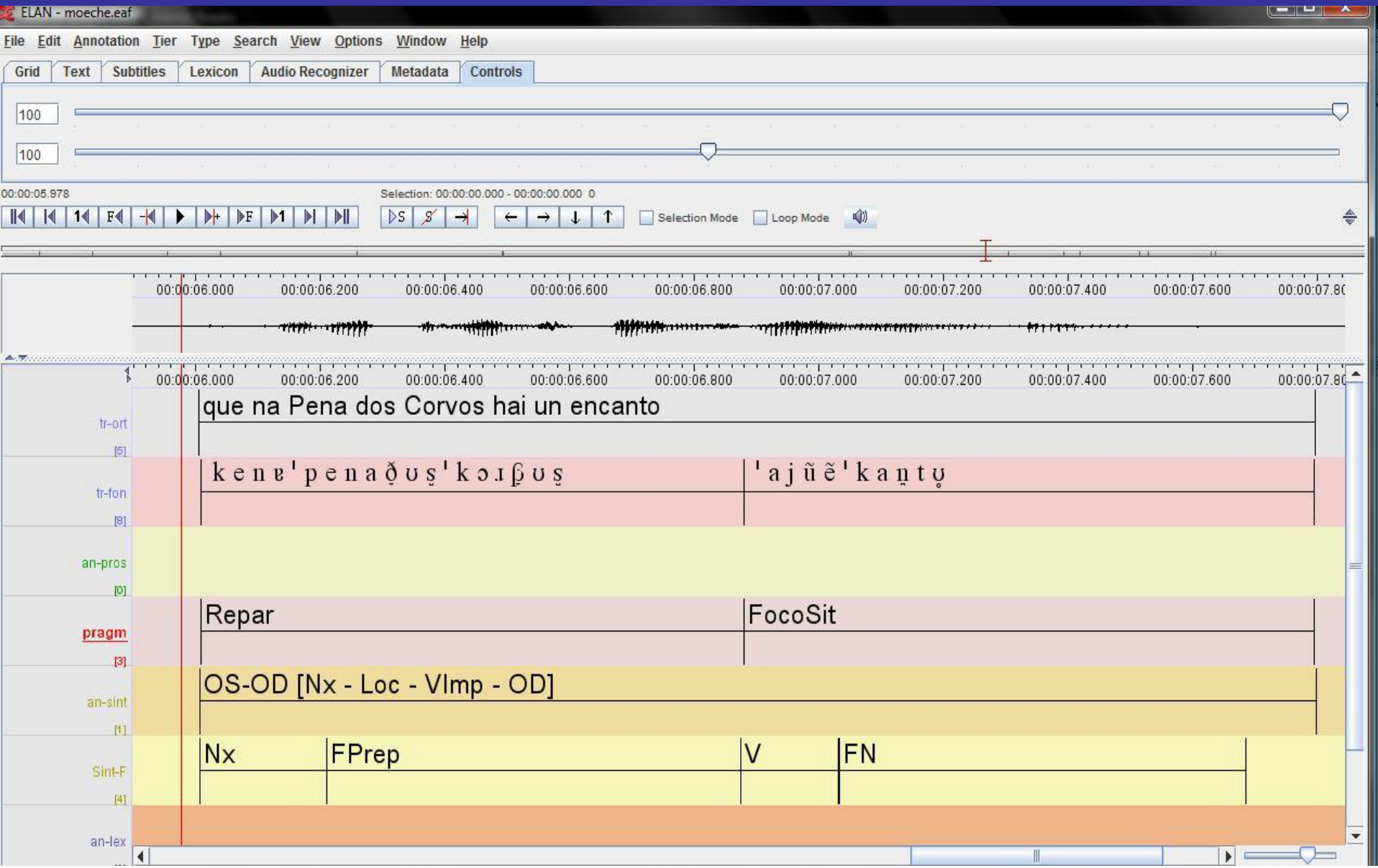
CORILGA: Aplicacións

- Lingüística baseada en corpus: estudos fonéticos, prosódicos, gramaticais, etc.
- Estudos lingüísticos interdisciplinares: relación entre niveis (fonética, prosodia, gramática, discurso, pragmática…)
- Desenvolvemento de tecnoloxías da fala
- Material para o ensino da lingua

ELAN (http://www.lat-mpi.eu/tools/elan/)



CORILGA


ELAN - moeche.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Audio Recognizer Metadata Controls
100
100
00:00:05.978
Selection: 00:00:00.000 - 00:00:00.000 0
Selection Mode
Loop Mode
00:00:06.000
00:00:06.200
00:00:06.400
00:00:06.600
00:00:06.800
00:00:07.000
00:00:07.200
00:00:07.400
00:00:07.600
tr-ort
que na Pena dos Corvos hai un encanto
tr-fon
k e n ɐ ' p e n a ð ʊ s ' k ɔ ɾ β ʊ s
' a j ũ ẽ ' k a n̪ t ʊ̯
an-pros
pragm
Repar
FocoSit
an-sint
OS-OD [Nx - Loc - VImp - OD]
Sint-F
Nx
FPrep
V
FN
an-lex

Cara ao futuro

- Corpus oral da variación lingüística en Galicia (variedades galegas e castelás, cambios de código e hibridación)
- ELAN:
  - posibilita o enlace co subcorpus galego de PRESEEA: PRESEGAL http://gramatica.usc.es/proxectos/presegal/
  - Posibilita o enlace co Corpus do Português Oral (CORP-ORAL) http://www.iltec.pt/spock/
- Dificultades?

Colaboración

- Informática:
  - mellorar a conexión ELAN – PRAAT
  - motor de busca
- Tecnoloxías e PLN:
  - recoñecedor-transcritor
  - ferramentas para a etiquetaxe: prosodia, gramática
- Estatística:
  - Tratamento estatístico de datos