

Corpus orales: ESLORA

Victoria Vázquez Rozas



<http://eslora.usc.es>



Grupo de investigación Gramática del español

Universidade de Santiago de Compostela

Proyectos de investigación

ESLORA FFI2010-17417

ESLORA2 FFI2014-52287-P

ESLORA+ PFFI2017-86379-P

Equipo actual del proyecto ESLORA

Mario Barcala

Montserrat Recalde

Marta Blanco

Raquel Rivas

Eva Domínguez

Guillermo Rojo

Alba Fernández

Paula Santalla

Francisco García

Victoria Vázquez

Sol López

Contenidos

1. El registro del habla
2. Construcción de un corpus oral
 - 2.1. Objetivos
 - 2.2. Diseño y elaboración
3. Decisiones de codificación
4. Explotación del corpus
 - 4.1. La aplicación de consulta
 - 4.2. Algunos ejemplos
5. Casos prácticos

2. Construcción de un corpus oral

2.1. Objetivos

2.2. Diseño y elaboración

Algunos corpus orales del español

Un corpus es un conjunto de (fragmentos de) textos naturales, almacenados en formato electrónico, representativos en su conjunto de una variedad lingüística, en alguno de sus componentes o en su totalidad, y reunidos con el propósito de facilitar su estudio científico (Rojo 2016: 285)

- Valesco 2.0
<http://valesco.es/?q=corpus>
- PRESEEA
<http://preseea.linguas.net/Corpus.aspx>
- Corpus Oral y Sonoro del Español Rural (COSER)
<http://corpusrural.org/coser/>
- C-Or-DiAL
http://lablita.dit.unifi.it/corpora/index_html/C-Or-DiAL/
- ESLORA
<http://eslora.usc.es/>

Objetivos


Objetivos

Aportar nuevos materiales de español oral

Comparar métodos de obtención de habla

Desarrollar herramientas de enriquecimiento y explotación de los datos

Uso habitual (IGE 2013)



Corpus: entrevistas Hablante: SCOM_M32_025_hab1 Papel: informante Sexo: ...

> ^ <ruido tipo="chasquido boca"/> le pedimos una <pausa/> que nos fa

> ^ que nos cont <pausa/> que nos <pausa/> pusiera en contacto con a

> y que y que nos ayudara <pausa/> a ver eeh <pausa/> Estambul <pausa/>

y	que	y	que	nos	ayudara	<pausa/>	a	ver	eeh	<pausa/>
C	C	C	C	PY1EPO	vs13s	ETQ_PAUSA	X	VNP	I	ETQ_PAUS
y	que	y	que	nos	ayudar	<pausa/>	a	ver	eeh	<pausa/>

< >

> ^ y y queríamos ver eeh <pausa/> Estambul <pausa/> por dentro <sile

> ^ y efectivamente <pausa/> nos nos <pausa/> presentó nos pusimos e Nasli <pausa_larga/>

lora.usc.es/search/context/d3fc26f8-bd06-4189-a87a-20e...a2c720a?index=3&number_search

Diseño y elaboración

Entrevistas semidirigidas

- 54 entrevistas (60 horas audio)
- Hablantes de Santiago de Compostela
- Metadatos
 - [cuestionario sociolingüístico](#)
 - [test inseguridad](#)
- Transcripción ortográfica alineada manualmente (*Transcriber*)

Conversaciones espontáneas

- 20 horas de grabación de audio
- Hablantes de Galicia (pero no solo)
- Metadatos
 - [fichas de hablante](#)
 - [fichas de evento comunicativo](#)
- Transcripción ortográfica alineada manualmente (*ELAN*)

V. 1.2.2 de nov. 2018

- 647.758 palabras / 776.260 elementos gramaticales
- 53 entrevistas y 3 conversaciones

Diseño y elaboración

Condiciones para la grabación de los intercambios

- Información previa sobre informantes (entrevistas)
- Condiciones de espacio y previsión de tiempo
- Preparación de los instrumentos de grabación
- Desarrollo de la grabación
- Reflexión metodológica permanente

Diseño y elaboración

Garantías éticas y legales para los informantes

- Consentimiento informado
 - Información sobre la finalidad del proyecto
 - Formularios de consentimiento ([entrev.](#), [previo](#) y [posterior](#))
- Compromiso de confidencialidad
 - Protección de datos personales (LOPD)
 - Garantía de anonimato en las transcripciones
- Control de uso y difusión de las grabaciones y las transcripciones

Diseño y elaboración

Anonimización de los materiales

La anonimización consiste en la modificación o eliminación de aquellos elementos que permitan identificar a los participantes o a cualquier otra persona que aparezca mencionada en la grabación.

- Identificadores directos: nombre, apodos
 - Identificadores indirectos: localizaciones geográficas
-
- Valoración de la conveniencia de regular el acceso

3. Decisiones de codificación

Representación de los datos

¿Qué información representamos?

- Dinámica de la toma de turno
- Intervenciones de los participantes
- Elementos no verbales (multimodalidad)

El detalle y las características dependen de

- La finalidad de los datos
- Los medios humanos y materiales disponibles
- La alineación de la transcripción con el sonido

Representación de los datos

Interjecciones y elementos “cuasi-léxicos”

1. lo malo que hay eh pue pue <pausa/> nos pues bueno nosotros con el repertorio que tenemos ya nos llega **ho** <pausa/> (SCOM_H31_046)
2. en esa excursión fue como muy divertido ¿ sabes ? en plan <pausa/> españoles se encuentran a un español <pausa/> famoso y vas en plan **oh** no sé qué ¿ sabes ? <pausa/> fue como muy <risa/> <pausa/> muy gracioso sí <pausa/> (SCOM_M13_008)

Representación de los datos

Variantes morfológicas

3. entonces es diferente <pausa/> en Cuba no tanto pero bueno <pausa/> Venezuela <pausa/> Colombia ya no te cuento y todos **eses** países sí sí <pausa/> (SCOM_M12_020)
4. y yo le dije a mi marido dije oy **hijño** no mejor es no ir porque pft (SCOM_M33_009)
5. y allá mi padre <ruido tipo="golpe"/> pero ¡ tú qué **hicistes** tal ! <pausa_larga/> es que eso lo tengo como si <pausa/> <risa/> mira como si lo estuviera viendo ahora <pausa_larga/> (SCOM_H21_053)

Codificación diferencial

I: en otras cosas / hay esos cambios ya más acelerados / bueno esas historias la primavera <risas=E> la sangre altera / patatín patatán pero / no <alargamiento> SCOM_H13_012

M1: yo ¿dónde me pongo?

M2: ahí [donde estás]

M3: [*inint*> ahí donde] estás // quieta

M1: ¡ay!

M3: o te mato

M1: no= [/ <@ @>]

M2: [<@ @> / <@ @>]

H3: [te voy a desenredar] ¿vale?

(SCOM_C_08)

Codificación diferencial

I: en otras cosas / hay esos cambios ya más acelerados / bueno esas historias la primavera <risas=E> la sangre altera / patatín patatán pero / no <alargamiento> SCOM_H13_012

M1: yo ¿dónde me pongo?

M2: ahí [donde estás]

M3: [<inint> ahí donde] estás // quieta

M1: ¡ay!

M3: o te mato

M1: no= [/ <@ @>]

M2: [<@ @> / <@ @>]

H3: [te voy a desenredar] ¿vale?

(SCOM_C_08)

Codificación integrada

C: he lla- he llamao a mi casa estaba mi abuelita y le digo *(ab)uelita, ¿está la mamá? diu¹ noo digo y el papá? diu noo*↑(RISAS) // *uelita², pues diles // QUE HE APROBADO*↑=

A: [EL CARNÉ DE LA AUTO]

C:=[EL CARNÉ DE COCHE]

C: y enseguida diu *¿el quéeee?* y digo *uelita el coche digoll ¿te acordarás? dice / ¿de quéé↑? i jo dic³ [ayy]=*

A: [aaah]

C:= *que no se acordaría que noo*

(Val.Es.Co 146 A1)

Codificación diferencial

"y <pausa/> bueno <ficticio>Laura</ficticio> y yo pa pa pa así riéndonos <cita_inicio/><lengua_inicio nombre="" gl""/>eh <pausa/> non sei de que se ríen total non sei <vacilación/> non entenden non sei que<lengua_fin/><cita_fin/> <pausa/> <palabra_cortada>bue</palabra_cortada> nosotros acabamos <pausa/> yo es que ya no aguantaba más <risa/> es que no aguantaba más <pausa_larga/>" (ESLORA SCOM_H21_039)

Corpus textuais (USC, 2019)

Codificación integrada

C: he lla- he llamao a mi casa estaba mi abuelita y le digo *(ab)uelita, ¿está la mamá? diu¹ noo digo y el papá? diu noo*↑(RISAS) // *uelita², pues diles // QUE HE APROBADO*↑=

A: [EL CARNÉ DE LA AUTO]

C:=[EL CARNÉ DE COCHE]

C: y enseguida diu *¿el quéeee?* y digo *uelita el coche digoll ¿te acordarás? dice / ¿de quéé↑? i jo dic³ [ayy]=*

A: [aaah]

C:= *que no se acordaría que noo*

(Val.Es.Co 146 A1)

Codificación diferencial

y <pausa/> bueno <ficticio>Laura</ficticio> y yo pa pa pa así riéndonos
<cita_inicio/><lengua_inicio nombre=""gl""/>eh <pausa/> non sei de que se
ríen total non sei <vacilación/> non entenden non sei
que<lengua_fin/><cita_fin/><pausa/>
<palabra_cortada>bue</palabra_cortada> nosotros acabamos <pausa/> yo
es que ya no aguantaba más <risa/> es que no aguantaba más
<pausa_larga/> (ESLORA SCOM_H21_039)

Corpus textuais (USC, 2019)

Etiquetas

Transcriber

```
<Event desc="alargamiento"  
type="pronounce"  
extent="instantaneous"/>
```

```
<Event desc="cita" type="lexical"  
extent="begin"/>
```

```
<Event desc="cita" type="lexical"  
extent="end"/>
```

```
<Event desc="palabra cortada"  
type="pronounce"  
extent="instantaneous"/>
```

```
<Event desc="risa=l" type="lexical"  
extent="instantaneous"/>
```

ELAN

=

```
<cita></cita>
```

-

```
<@ @/>
```

ESLORA

```
<alargamiento></alargamie  
nto>
```

```
<cita_inicio></cita_fin>
```

```
<palabra_cortada></palabr  
a_cortada>
```

```
<risa/>
```

¿?	Enunciado interrogativo
¡!	Enunciado exclamativo
~Nombre	Nombre ficticio
<alargamiento></alargamiento>	Aumento de cantidad en algún sonido de la palabra marcada
<cita_inicio/> <cita_fin/>	Inicio y fin de fragmento en estilo directo
<énfasis_inicio/> <énfasis_fin/>	Segmento pronunciado con especial intensidad
<ininteligible>	Sustituye un fragmento no comprensible y por tanto no transcrito
<lengua_inicio nombre="xx"/> <lengua_fin/>	Fragmento en una lengua distinta del español (gl: gallego, en: inglés, pt: portugués, it: italiano, etc.)
<palabra_cortada> </palabra_cortada>	Fragmento de una palabra
<pausa/>	Pausa breve
<pausa_larga/>	Pausa más larga pero inferior a un segundo
<risa/>	Risa de un/a hablante
<risa_inicio/> <risa_fin/>	Fragmento pronunciado entre risas
<ruido tipo="">	Entre las comillas se especifica el tipo de ruido
<sic_inicio/> <sic_fin/>	Lapsus de dicción que no debe confundirse con un error de transcripción
<sigla_inicio/> <sigla_fin/>	Sigla
<silencio>	Pausa de más de un segundo
<transcripción_dudosa_inicio/> </transcripción_dudosa_fin/>	La transcripción es problemática
<vacilación>	Sustituye fragmentos similares a palabras cortadas pero imposibles de transcribir

Lematización y etiquetación morfosintáctica

```
<fragmento hablante="habl" comienzo="108219.0" fin="115088.0">
  <expresión>¿sabes? no <risa/> <pausa/> yo no quiero saber <alargamiento>si</alargamiento>
  una cosa es un adverbio otra cosa <risa_inicio/>es un pronombre <risa_fin/> ¿sabes? <pausa/>
  que no te van a valer <alargamiento>mucho</alargamiento> <pausa/></expresión>
  <análisis>
    <análisis_unidad>
      <unidad>¿</unidad>
      <constituyente>
        <forma>¿</forma>
        <etiqueta>Q</etiqueta>
        <lema>¿</lema>
      </constituyente>
    </análisis_unidad>
    <análisis_unidad>
      <unidad>sabes</unidad>
      <constituyente>
        <forma>sabes</forma>
        <etiqueta>VIP2S</etiqueta>
        <lema>saber</lema>
      </constituyente>
    </análisis_unidad>
```

4. Explotación del corpus

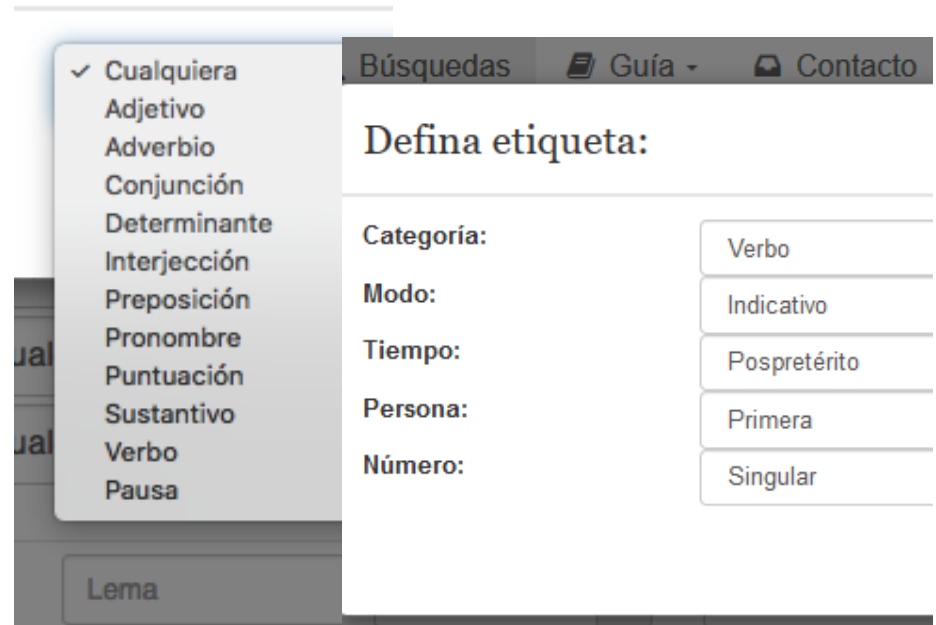
4.1. La aplicación de consulta

4.2. Algunos ejemplos

Aplicación de consulta

Criterios de búsqueda

- **Tipo de discurso**
 - Entrevista
 - Conversación
- **Tipo de búsqueda**
 - Palabra ortográfica
 - Elemento gramatical
 - Lema
 - Categoría gramatical
- **Datos de los hablantes**
 - Edad
 - Papel comunicativo
 - Género
 - Nivel educativo



- **Contexto oral**
 - cita
 - risas
 - alargamiento

Aplicación de consulta

<http://eslora.usc.es/>

ESLORA [Información](#) [Búsquedas](#) [Guía](#) [Contacto](#) [Descargas](#) [Equipo](#) [Acerca de](#)

Búsqueda

Corpus:

Tipo:

Sensibilidad

Acentos:

Mayúsculas:

Resultado

Tipo:

Ordenación:

Tamaño página:

Filtros

Edad:

Papel:

Sexo:

Desde:

Hasta:

Estudios:

Buscar en:

Texto

[Volver](#) [Limpiar](#) [Buscar](#)

Corpus textuais (USC, 2019)

Aplicación de consulta

eslora.usc.es/data 67% corpus eslora

ESLORA Información Búsquedas Guía · Contacto Descargas Equipo Acerca de

En su versión actual, 1.2 de octubre de 2018 , el corpus disponible para consulta consta de 56 documentos que incluyen 647.757 palabras ortográficas (776.250 elementos gramaticales), que se distribuyen de la siguiente forma:

Palabras ortográficas

Grupo de edad		
	Palabras ortográficas	Documentos
19-34	214.614	44
35-54	199.034	27
>54	224.739	19
Desconocido	794	18

Papel		
	Palabras ortográficas	Documentos
Audiencia	2.196	23
Entrevistador	72.531	53
Informante	573.030	56

Sexo		
	Palabras ortográficas	Documentos
Hombre	275.966	34
Mujer	371.791	55

Nivel de estudios		
	Palabras ortográficas	Documentos
Universitarios	272.955	55
Medios	182.665	20
Primarios	191.342	17
Desconocidos	795	18

Corpus		
	Palabras ortográficas	Documentos
Entrevistas	639.181	53
Conversaciones	8.576	3

5. Casos prácticos

Tipo de búsqueda: palabras ortográficas

Texto

super*!super!supera*

Resultados 1 a 50 de 100

Anterior 1 2 Siguiente

experiencia que tuve <pausa/> mm conmigo se portaron	superbién <silencio/>
que es una zona	superbonita <pausa/> que llegas <pausa/> llegas a una zona <p
y estás allí <pausa/> un agua cristalina	superbonito <pausa_larga/>
estaba yo contemplando el río no sé qué no sé cuantos	superbonito no sé qué <pausa/> y viendo el puente <pausa/> y l
<pausa/> en clase era ¡ fuera ! <pausa/> y un día me ve	supercallada supertal super <pausa/> y y me <pausa/> me dijo p
me acuerdo de mm de eso porque yo <pausa/> como era	supercallejera que a mí me encantaba estar en la calle <pausa_
a un restaurante	supercaro <pausa/> no es eeh <pausa/> entonces nunca tuve un
a/> el auditorio para pasear cerquita <pausa/> está todo	supercerca eh andando <pausa/> está todo más cerca que en co

Tipo de búsqueda: elementos gramaticales

Elementos gramaticales

Elemento gramatical: Palab. ortográfica

[Volver](#)

Hay 90 / 776.260 coincidencias (116/millón) en 32 / 56 documentos.

Elementos gramaticales

ifíc|*ifíq* Lema Palab. ortográfica

[Volver](#) [Descargar](#)

Resultados 1 a 50 de 92

[Anterior](#) **1** [2](#) [Siguiente](#)

Benjamín Julián <pausa/> y el beato Luciano <pausa/> que fue	beatificado <pausa/> el el hace hace un mes <pausa/> es de estos cuatrocientos nove
aquí en la aldea aquí <pausa/> a nosotros nos tienen	calificado como mmm <pausa/> la la ciudad sin ley <pausa_larga/>
e ir a Correos <pausa/> a	certificar <pausa/> o poner un giro postal <pausa_larga/>
mmm <pausa/> yo <pausa/> te mmm <pausa/>	certifico <pausa/> que vienen a prácticas <pausa/> por lo tanto <pausa/> no tienes nir
mmm tú exactamente	clasificar todo aquello <pausa/> que estás viviendo <pausa/> pero vo <pausa/> tengo

Tipo de búsqueda: elementos gramaticales

Elementos gramaticales

!*da!*das VPF* ? Lema Palab. ortográfica

Volver Descargar

Resultados 1 a 74 de 74

eeh botella que yo vea que está	abierta <silencio/>
<ininteligible/> yo la deajo	abierta <ruido tipo="puerta"/> en media hora nos va <silencio/>
eeh zona así toda	abierta pero <pausa/> con tapado como si fuera trincheras y todo eso <pausa/>
ntemente sí pero <pausa/> es decir y luego <pausa/> es decir no sé si ir con una cervecita <pausa/>	abierta así <pausa/> bebiendo <pausa/> incluye beber en la vía pública <pausa/>

Tipo de búsqueda: elementos gramaticales próximos

Elementos gramaticales

Elemento gramatical	V*	?	ser	Palab. ortográfica	1
Elemento gramatical	A* VP*	?	Lema	Palab. ortográfica	

Volver Descargar Limpiar Bus

Resultados 1 a 50 de 1.445

Anterior **1** 2 3 4 5 ... 28 29 Siguiente

1 Ir a la página

porque él mi dijo si mmm tenía <pausa/> si	era alérgica ¿ no sabes ? cuando te preguntan a mmm a algo y ta
¿ era alquila ? ¿	era alquilado ?
y el jefe a lo mejor le reñía <pausa/> el jefe daba la vuel	era alto era un señor muy serio <pausa/> y daba la vuelta y se po
hacías aunque yo	era auxiliar <pausa/> hacías la función de de enfermera ¿ no ? <
y aún eh yo	era auxiliar <pausa/> aún trabajé en este hospital unos meses de

Tipo de búsqueda: elementos gramaticales próximos

Elementos gramaticales

Elemento gramatical	V*	?	ser	Palab. ortográfica	≤ 2
Elemento gramatical	A* VP*	?	Lema	Palab. ortográfica	

Volver Descargar Limpiar

Resultados 1 a 50 de 2.801

Anterior **1** 2 3 4 5 ... 56 57 Siguiente

1 Ir a la página

llegaba el suelo <pausa/> desde las ventanas hasta la mitad y después	era todo abierto hacia la hacia hasta arriba hasta el techo <pausa_larga/>
gún algún pianista algún músico <pausa/> comiendo <pausa/> mi padre	era muy aficionado a <pausa/> mis padres se fueron siempre muy abiertos muy <pausa/>
mi padre	era muy aficionado y mi madre tocó <pausa/> tenía la carrera entera de música <pausa_la
ues <pausa/> eeh <pausa/> y entonces parece que el el el cardenal no	era muy aficionado a tener servicio y gente allí en palacio <pausa/> <ininteligible/> simpler
a/> antes era de <pausa/> subiendo a la mano izquierda <pausa/> antes	era todo ajardinado <pausa/> y jugábamos allí y me acuerdo una vez de pequeño <pausa/
porque mmm <pausa/>	era muy alegre muy mmm <pausa/> sí <pausa/> además después ¿ qué te voy a decir ? <pausa/
porque él mi dijo si mmm tenía <pausa/> si	era alérgica ¿ no sabes ? cuando te preguntan a mmm a algo y tal yo le dije bueno <pausa/
¿ era alquila ? ¿	era alquilado ?
h/#	era un alto cargo en una <pausa/> en Madrid <pausa_larga/>