

# Una metodología semiautomática de anotación de entidades nombradas para la creación de un *gold standard*

Susana Sotelo<sup>†</sup> Pablo Gamallo<sup>†</sup> Álvaro Iriarte<sup>‡</sup>

<sup>†</sup> CiTIUS – Universidade de Santiago

<sup>‡</sup> CEHUM – Universidade do Minho



Universidade do Minho

V Congreso de la Soc. Internacional de Humanidades Digitales Hispánicas

Santiago de Compostela, 4-8 de octubre de 2021

# ÍNDICE

## 1 CONTEXTO Y OBJETIVOS

## 2 CORPUS

Descripción

Proceso de codificación

Proceso de anotación de entidades nombradas

## 3 EXPLOTACIÓN



Rede Galabra

# CONTEXTO

## Proyectos de investigación:

- *Discursos, imágenes y prácticas culturales de Santiago de Compostela como meta del Camino* (FFI2012-35521)
- *Narrativas, usos y consumos de visitantes como aliados o amenazas para el bienestar de la comunidad local: el caso de Santiago de Compostela* (FFI2017-88196-R)

## Objetivo:

- Identificar las prácticas efectivas de los visitantes de Santiago de Compostela y cruzarlas con aquellas presentes en productos culturales y literarios
- Analizar el impacto de dichas prácticas sobre la comunidad local

## Recursos desarrollados:

- 1 Corpus literario: catálogo de 546 libros con discursos y prácticas culturales relativas a Santiago de Compostela o al Camino de Santiago
- 2 Redes sociales: en torno a 350.000 *tweets* mencionando el Camino de Santiago
- 3 Base de datos de encuestas a visitantes de Santiago de Compostela y transcripción de las entrevistas telefónicas a parte de los encuestados

# CONTEXTO

## Proyectos de investigación:

- *Discursos, imágenes y prácticas culturales de Santiago de Compostela como meta del Camino* (FFI2012-35521)
- *Narrativas, usos y consumos de visitantes como aliados o amenazas para el bienestar de la comunidad local: el caso de Santiago de Compostela* (FFI2017-88196-R)

## Objetivo:

- Identificar las prácticas efectivas de los visitantes de Santiago de Compostela y cruzarlas con aquellas presentes en productos culturales y literarios
- Analizar el impacto de dichas prácticas sobre la comunidad local

## Recursos desarrollados:

- 1 Corpus literario: catálogo de 546 libros con discursos y prácticas culturales relativas a Santiago de Compostela o al Camino de Santiago
- 2 Redes sociales: en torno a 350.000 *tweets* mencionando el Camino de Santiago
- 3 Base de datos de encuestas a visitantes de Santiago de Compostela y transcripción de las entrevistas telefónicas a parte de los encuestados

# CONTEXTO

## Proyectos de investigación:

- *Discursos, imágenes y prácticas culturales de Santiago de Compostela como meta del Camino* (FFI2012-35521)
- *Narrativas, usos y consumos de visitantes como aliados o amenazas para el bienestar de la comunidad local: el caso de Santiago de Compostela* (FFI2017-88196-R)

## Objetivo:

- Identificar las prácticas efectivas de los visitantes de Santiago de Compostela y cruzarlas con aquellas presentes en productos culturales y literarios
- Analizar el impacto de dichas prácticas sobre la comunidad local

## Recursos desarrollados:

- 1 Corpus literario: catálogo de 546 libros con discursos y prácticas culturales relativas a Santiago de Compostela o al Camino de Santiago
- 2 Redes sociales: en torno a 350.000 *tweets* mencionando el Camino de Santiago
- 3 Base de datos de encuestas a visitantes de Santiago de Compostela y transcripción de las entrevistas telefónicas a parte de los encuestados

# DESCRIPCIÓN DEL CORPUS

- 1 Base de datos con 2080 encuestas a visitantes de la ciudad de Santiago
  - Muestreo no probabilístico accidental (por lugar de residencia)
  - Encuestas a pie de calle en puntos de interés turístico, zonas de esparcimiento y terminales de transporte, entre 2013 y 2014
- 2 Corpus de 252 entrevistas telefónicas
  - Semiestructuradas
  - En tres idiomas: español, gallego y portugués

## Composición:

Por origen del informante:

	Encuestas	Porcentaje	Entrevistas	Porcentaje
Galicia	398	19,13 %	47	18,65 %
España	878	42,21 %	92	36,51 %
Portugal	408	19,62 %	54	21,43 %
Brasil	396	19,04 %	59	23,41 %
<b>Total</b>	<b>2080</b>	<b>100,00 %</b>	<b>252</b>	<b>100,00 %</b>

Por lengua y número de palabras:

Lengua	Palabras	Porcentaje
gallego	177.999	12,01 %
español	655.112	44,22 %
portugués	648.404	43,77 %

# DESCRIPCIÓN DEL CORPUS

- 1 Base de datos con 2080 encuestas a visitantes de la ciudad de Santiago
  - Muestreo no probabilístico accidental (por lugar de residencia)
  - Encuestas a pie de calle en puntos de interés turístico, zonas de esparcimiento y terminales de transporte, entre 2013 y 2014
- 2 Corpus de 252 entrevistas telefónicas
  - Semiestructuradas
  - En tres idiomas: español, gallego y portugués

## Composición:

Por origen del informante:

	Encuestas	Porcentaje	Entrevistas	Porcentaje
Galicia	398	19,13 %	47	18,65 %
España	878	42,21 %	92	36,51 %
Portugal	408	19,62 %	54	21,43 %
Brasil	396	19,04 %	59	23,41 %
<b>Total</b>	<b>2080</b>	<b>100,00 %</b>	<b>252</b>	<b>100,00 %</b>

Por lengua y número de palabras:

Lengua	Palabras	Porcentaje
gallego	177.999	12,01 %
español	655.112	44,22 %
portugués	648.404	43,77 %

# DESCRIPCIÓN DEL CORPUS

- 1 Base de datos con 2080 encuestas a visitantes de la ciudad de Santiago
  - Muestreo no probabilístico accidental (por lugar de residencia)
  - Encuestas a pie de calle en puntos de interés turístico, zonas de esparcimiento y terminales de transporte, entre 2013 y 2014
- 2 Corpus de 252 entrevistas telefónicas
  - Semiestructuradas
  - En tres idiomas: español, gallego y portugués

## Composición:

Por origen del informante:

	Encuestas	Porcentaje	Entrevistas	Porcentaje
Galicia	398	19,13 %	47	18,65 %
España	878	42,21 %	92	36,51 %
Portugal	408	19,62 %	54	21,43 %
Brasil	396	19,04 %	59	23,41 %
<b>Total</b>	<b>2080</b>	<b>100,00 %</b>	<b>252</b>	<b>100,00 %</b>

Por lengua y número de palabras:

Lengua	Palabras	Porcentaje
gallego	177.999	12,01 %
español	655.112	44,22 %
portugués	648.404	43,77 %



# PROCESO DE CODIFICACIÓN

## Fases de codificación

- 1 Transcripción a partir de grabaciones
  - Sin representación de mecanismos propios de la oralidad (vacilaciones, interrupciones, etc.)
  - Realizada por un equipo heterogéneo de personas
- 2 Corrección y armonización de los textos
  - Corrección y normalización ortográfica
  - Fijación de omisiones y errores por distancia cultural del transcriptor

Algunos ejemplos: *fevereiro* en lugar de *Cebreiro*, *fui no ferry* por *Finisterre* o *farol* por *Ferrol*
- 3 Codificación en XML con esquema propio:  
<https://galabra.linguarum.net/corpus/schemadoc/>
  - Sección con metadatos
  - Identificación de turnos (entrevistador/entrevistado) y lengua de cada turno

Rede Galabra

# PROCESO DE CODIFICACIÓN

## Fases de codificación

- 1 Transcripción a partir de grabaciones
  - Sin representación de mecanismos propios de la oralidad (vacilaciones, interrupciones, etc.)
  - Realizada por un equipo heterogéneo de personas
- 2 Corrección y armonización de los textos
  - Corrección y normalización ortográfica
  - Fijación de omisiones y errores por distancia cultural del transcriptor  
Algunos ejemplos: *fevereiro* en lugar de *Cebreiro*, *fui no ferry* por *Finisterre* o *farol* por *Ferrol*
- 3 Codificación en XML con esquema propio:  
<https://galabra.linguarum.net/corpus/schemadoc/>
  - Sección con metadatos
  - Identificación de turnos (entrevistador/entrevistado) y lengua de cada turno

Rede Galabra

# PROCESO DE CODIFICACIÓN

## Fases de codificación

- 1 Transcripción a partir de grabaciones
  - Sin representación de mecanismos propios de la oralidad (vacilaciones, interrupciones, etc.)
  - Realizada por un equipo heterogéneo de personas
- 2 Corrección y armonización de los textos
  - Corrección y normalización ortográfica
  - Fijación de omisiones y errores por distancia cultural del transcriptor  
Algunos ejemplos: *fevereiro* en lugar de *Cebreiro*, *fui no ferry* por *Finisterre* o *farol* por *Ferrol*
- 3 Codificación en XML con esquema propio:  
<https://galabra.linguarum.net/corpus/schemadoc/>
  - Sección con metadatos
  - Identificación de turnos (entrevistador/entrevistado) y lengua de cada turno

Rede Galabra

# PROCESO DE CODIFICACIÓN

```

1 <?xml version="1.0" encoding="utf-8" standalone="no" ?>
2 <transcription xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="transcription.xsd" >
3 <metadata >
4   <participants >
5     <speaker id="119" >
6       <age value="2" />
7       <gender value="H" />
8       <grade value="7" />
9       <pilgrim value="1" />
10      <origin value="PT" />
11    </speaker >
12    <researcher id="5" name="Ana Rita" />
13  </participants >
14  <source phase="1" >
15    <text >
16      <file name="Fator09_FASE1_PT_119_CristianoRodrigues_20-02-2014.doc" type="doc" />
17    </text >
18    <audio duration="2701" >
19      <file name="00_PT_fase1_119.mp3" type="mp3" />
20    </audio >
21  </source >
22 </metadata >
23 <data >
24 <q lang="pt" >Bom dia. O meu nome é Ana Rita. Ligo-lhe desde a Universidade de Santiago. Esteve a trocar uns emails comigo.</q>
25 <a lang="pt" >Sim, estou recordado. Por causa de um estudo?</a >
26 <q lang="pt" >Exatamente, sobre Santiago. Gostaríamos de saber um pouco mais de aquilo que disse ao meu colega.</q >
27 [...]
```

# PROCESO DE ANOTACIÓN

## Anotación NER

Tarea del Procesamiento del Lenguaje Natural que tiene por objeto la detección y clasificación de entidades nombradas en un corpus textual.

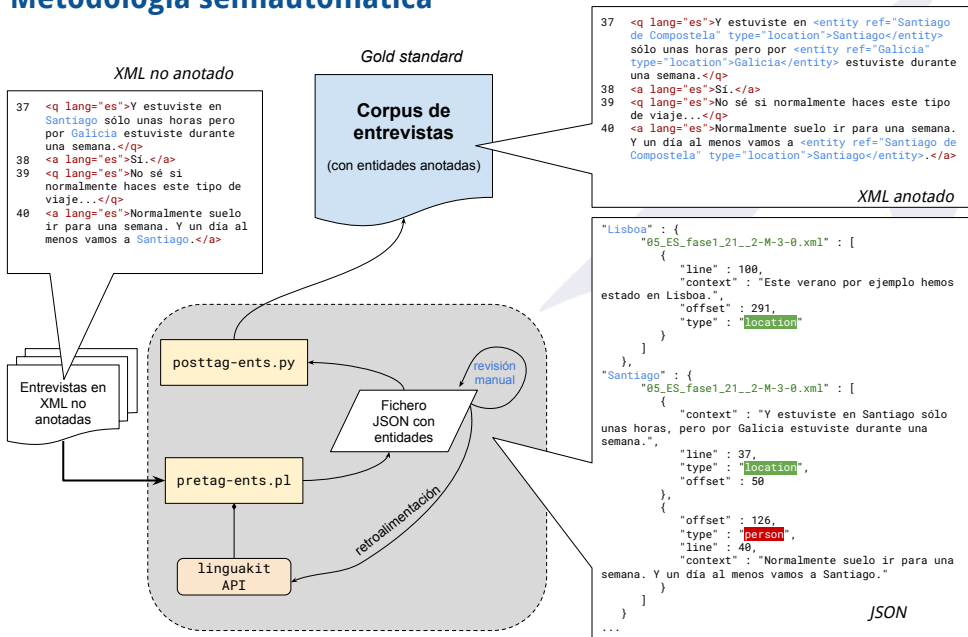
### ① Herramienta: Linguakit → <https://github.com/citiususc/Linguakit>

- Entidades *numex* (de base numérica) y *enamex* (nombres propios)
- Método de supervisión distante para la clasificación en 4 tipos: *persona*, *organización*, *localización* y *miscelánea*
- Soporte para todas las lenguas del proyecto: español, gallego y portugués
- Software libre

### ② Resultados

	español		gallego		portugués		Total
location	18323	85.79 %	5921	83.36 %	12183	85.71 %	36427
person	1207	5.63 %	370	5.21 %	1063	7.48 %	2640
org	436	2.05 %	200	2.81 %	399	2.81 %	1035
misc	1393	6.53 %	612	8.62 %	568	4.00 %	2573
<b>Total</b>	<b>21359</b>	<b>100 %</b>	<b>7103</b>	<b>100 %</b>	<b>14213</b>	<b>100 %</b>	<b>42675</b>

# Metodología semiautomática



XML no anotado

```

37 <q lang="es">Y estuviste en
    Santiago sólo unas horas pero
    por Galicia estuviste durante
    una semana.</q>
38 <a lang="es">Si.</a>
39 <q lang="es">No sé si
    normalmente haces este tipo de
    viaje...</q>
40 <a lang="es">Normalmente suelo
    ir para una semana. Y un día al
    menos vamos a Santiago.</a>
    
```

Gold standard

**Corpus de entrevistas**  
(con entidades anotadas)

```

37 <q lang="es">Y estuviste en <entity ref="Santiago
    de Compostela" type="location">Santiago</entity>
    sólo unas horas pero por <entity ref="Galicia"
    type="location">Galicia</entity> estuviste durante
    una semana.</q>
38 <a lang="es">Si.</a>
39 <q lang="es">No sé si normalmente haces este tipo
    de viaje...</q>
40 <a lang="es">Normalmente suelo ir para una semana.
    Y un día al menos vamos a <entity ref="Santiago de
    Compostela" type="location">Santiago</entity>.</a>
    
```

XML anotado

```

"lisboa" : {
  "05_ES_fase1_21_2-M-3-0.xml" : [
    {
      "line" : 100,
      "context" : "Este verano por ejemplo hemos
estado en Lisboa.",
      "offset" : 291,
      "type" : "location"
    }
  ]
},
"santiago" : {
  "05_ES_fase1_21_2-M-3-0.xml" : [
    {
      "context" : "Y estuviste en Santiago sólo
unas horas, pero por Galicia estuviste durante una
semana.",
      "line" : 37,
      "type" : "location",
      "offset" : 50
    },
    {
      "offset" : 126,
      "type" : "person",
      "line" : 40,
      "context" : "Normalmente suelo ir para una
semana. Y un día al menos vamos a Santiago."
    }
  ]
}
...
    
```

JSON

# PROCESO DE ANOTACIÓN

```

1 <?xml version="1.0" encoding="utf-8" standalone="no" ?>
2 <transcription xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation="transcription.xsd" >
3 <metadata >
4   <participants >
5     <speaker id="119" >
6       <age value="2" />
7       <gender value="H" />
8       <grade value="7" />
9       <pilgrim value="1" />
10      <origin value="PT" />
11    </speaker >
12    <researcher id="5" name="Ana Rita" />
13  </participants >
14  <source phase="1" >
15    <text >
16      <file name="Fator09_FASE1_PT_119_CristianoRodrigues_20-02-2014.doc" type="doc" />
17    </text >
18    <audio duration="2701" >
19      <file name="00_PT_fase1_119.mp3" type="mp3" />
20    </audio >
21  </source >
22 </metadata >
23 <data >
24 <a lang="pt" >Sim, estou recordado. Por causa de um estudo?</a >
25 <q lang="pt" >Bom dia. O meu nome é <entity ref="Ana Rita" type="person" >Ana Rita</entity >. Ligo-lhe desde a <entity ref="
  Universidade de Santiago" type="org" >Universidade de Santiago</entity >. Esteve a trocar uns emails comigo.</q >
26 <q lang="pt" >Exatamente, sobre <entity ref="Santiago de Compostela" type="location" >Santiago</entity >. Gostaríamos de saber um pouco
  mais de aquilo que disse ao meu colega.</q >
27 [...]
```

# EXPLOTACIÓN Y LÍNEAS DE TRABAJO

## Explotación

- 1 Análisis cuantitativos y cualitativos
  - Corpus disponible para los investigadores del proyecto
  - Formatos HTML, XML y TXT
- 2 Entrenamiento de un NER basado en aprendizaje automático

## Líneas de trabajo futuro

- Anonimización de datos personales
- Publicación bajo licencia libre
- Construcción de mapas de densidad de consumos y prácticas de visitantes

### Formulário de descarregamento do corpus de transcrições

Escolha as opções que você precisar para selecionar os ficheiros do corpus.

[Marcar todas](#)

<p>P6.1 Idade</p> <p><input type="checkbox"/> Sem dados</p> <p><input type="checkbox"/> &lt; 30</p> <p><input type="checkbox"/> 30 &lt;= x &lt; 50</p> <p><input type="checkbox"/> 50 &lt;= x &lt; 65</p> <p><input type="checkbox"/> &gt;= 65</p>	<p>P6.2 Sexo</p> <p><input type="checkbox"/> Mulheres</p> <p><input type="checkbox"/> Homens</p>	<p>P01 País</p> <p><input checked="" type="checkbox"/> Brasil</p> <p><input checked="" type="checkbox"/> Portugal</p> <p><input type="checkbox"/> Galiza</p> <p><input type="checkbox"/> Espanha</p>
<p>P6.3 Nivel de estudos</p> <p><input type="checkbox"/> Sem dados</p> <p><input type="checkbox"/> Sem estudos / básicos</p> <p><input type="checkbox"/> Primários / ESO / 6º ano</p> <p><input type="checkbox"/> Secundários / bacharelato / 12º ano</p> <p><input type="checkbox"/> Formação profissional (médio)</p> <p><input type="checkbox"/> Formação profissional (superior) / mestria</p> <p><input type="checkbox"/> Universitários médios / grado</p> <p><input type="checkbox"/> Universitários superiores / posgraduação</p>	<p>P121 É peregrino?</p> <p><input checked="" type="checkbox"/> Não</p> <p><input type="checkbox"/> Sim</p>	

Correio electrónico

Formato final 
 Completo
 Entidades originais

Só respostas
  Entidades normalizadas

[Descarregar 80](#)



# Gracias

Contacto: <[susana.sotelo.docio@usc.gal](mailto:susana.sotelo.docio@usc.gal)>

Rede Galabra