

Herramientas y recursos lingüísticos disponibles en el seno del grupo *Gramática do Español*

Coordinadora: Victoria Vázquez Rozas

Facultade de Filoloxía
Universidade de Santiago de Compostela

Taller ReTeLe

Salamanca, 13 de septiembre de 2016

Contenidos

1. Recursos
2. Herramientas

Recursos

Base de datos sintácticos (BDS)

- ▶ Corpus de 158.000 cláusulas anotadas sintácticamente.
- ▶ Monolingüe: español

Recursos

Base de datos de morfología del español (BDME)

- ▶ Red léxica donde los términos se relacionan por origen y base.
- ▶ Lenguas: español y latín

Recursos

Corpus Eslora

- ▶ Corpus oral anotado con 250K palabras
- ▶ Monolingüe: español

Recursos

Corpus CSA

- ▶ Corpus sintáctico anotado con información constitutivo-funcional (30K palabras).
- ▶ Monolingüe: español
- ▶ Licencia LGPL

Herramientas PLN

Proyecto de software libre **Linguakit**

- ▶ Análisis: PoS tagging, NERC, parsing de dependencias,
- ▶ Extracción: términos, relaciones semánticas, conceptos, opinión
- ▶ Resumidor, corrector/evaluador gramatical (gallego)
- ▶ Para 4 lenguas: castellano, gallego, inglés, portugués
- ▶ Licencia GPL (proyecto github)
- ▶ Demo (250K consultas por mes).

Léxico distribucional diacrónico

Buscador diacrónico

- ▶ Sirve para realizar estudios lexicológicos sobre el cambio semántico
- ▶ Dada una palabra, busca las palabras más similares distribucionalmente durante un período de tiempo.
- ▶ Demo.

Webs del grupo

- ▶ `http://gramatica.usc.es`
- ▶ `http://gramatica.usc.es/pln`
- ▶ `http://gramatica.usc.es/~gamallo`

Corpus: BDS

- Linguality type: **molingual**
- Language: **Spanish**
- Text format: **database mysql**
- Size
 - Number of sentences: **158K clauses**
 - Number of words: -
- Character encoding: -

Corpus

- Classification
 - Text type: **written text, narrative**
 - Register
- Annotation: **deep syntactic annotation**
- Creation
 - Creation mode: **manual**
- Version:
- Documentation:
<http://www.bds.usc.es>

Distribution

- Availability: **web**
- License
 - License: **<http://www.bds.usc.es/licencia.html>**
 - Restrictions
 - Download location: -
 - Distribution Access/Medium

Contact Person: **Guillermo Rojo**

Resource Creation

- Resource creator: **GE group**
- Funding project: **Corpus del español sintácticamente anotado y revisado CESAR+**

Spanish Morphology DataBase (BDME)

- Resource type: **morphological**
- Linguality type: **bilingual**
- Languages: **Spanish, Latin**
- Text format: **database**
- Size
 - Number of entries: -
- Character encoding: -

Lexical Conceptual Resource

- Domains: -
- Encoding
 - Encoding level: -
 - Linguistic information: **etimological and morphological links**
- Creation mode: -
- Version: -

Documentation:

<http://bdme.usc.es/cms/>

Distribution

- Availability: **web**
- License
 - License: -
 - Restrictions: -
 - Download location: -
 - Distribution Access/Medium: -

Contact Person

Resource Creation

- Resource creator: **María José Espinheira, Jesús Pena**

Funding project: **Estudio morfogenético del léxico español. Ref: FFI2012-38550.**

Corpus: ESLORA

- Linguality type: **Monolingual**
- Language: **Spanish**
- Text format: **aligned XML (.trs .eaf)**
- Size
 - Number of sentences
 - Number of words: **212702 words (249677 grammatical forms) (version 0.8.3 2015)***
- Character encoding: **UTF-8**

Corpus ESLORA

- Classification
 - Text type: oral
 - Register: semi-directed interviews and spontaneous conversations
- Annotation: POS and morphologically tagged
- Creation
 - Creation mode: manual and computational
- Version 0.8.3 (beta)
- Documentation

Distribution

- Availability: [free search online](#)
- License*
 - License Restrictions
 - Download location
 - Distribution Access/Medium

Contact Person [Victoria Vázquez Rozas](#)

Resource Creation

- Resource creator: [ESLORA/PRESEGAL team, Spanish Grammar Research Group \(USC\)](#)
 - **Search engine:** [Mario Barcala and team](#)
- Funding project: [ESLORA \(FFI2010-17417\)](#) and [ESLARA2 \(FFI2014-52287-P\)](#)

Corpus: CSA

- Linguality type: **molingual**
- Language: **Spanish**
- Text format: **plain text**
- Size
 - Number of sentences: **1354 (of which 1000 analysed)**
 - Number of words: **27827**
- Character encoding: **UTF8**

Corpus

- Classification
 - Text type: written text, narrative
 - Register
- Annotation: **deep syntactic annotation**
- Creation
 - Creation mode: **manual**
- Version: **1 beta**
- Documentation:
<http://gramatica.usc.es/wiki/drasae>

Distribution

- Availability: free resource
- License
 - License: Lesser General Public License for Linguistic Resources
 - Restrictions
 - Download location: <http://galvan.usc.es/drasae>
 - Distribution Access/Medium

Contact Person: M.^a Paula Santalla del Río

Resource Creation

- Resource creator: M.^a Paula Santalla del Río et al.
- Funding project: DRASAE,
<http://gramatica.usc.es/proyectos/drasae/?lang=gl&id=descripcion>

Additional observations: open for collaboration in maintenance and development