DH@Madrid Summer School 2016

Análisis computacional del soneto del Siglo de Oro - Parte 1

Borja Navarro Colorado - borja@dlsi.ua.es

European Research Council

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos

Digital Scholarly Editions Initial Training Network

Madrid, 1 de julio de 2016

Motivación

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE iNNOVACIÓN EN HUMANIDADES DIGITALES

- Visión panorámica del proceso de análisis computacional de un fenómeno literario.
  - Aplicación de la tecnología digital a un caso concreto.
  - Análisis a gran escala.

erc European Research Council

Objeto de estudio

- Aspectos métricos y semánticos (temático) del sonetos del Siglo de Oro.
  - S. XVI-XVII: de Garcilaso de la Vega a Sor Juana Inés de la Cruz.
  - Relevancia histórica del soneto, gran cantidad de autores.

Navarro Colorado (2015) “A computational linguistic approach...”

Índice

- Primera parte:
  - Compilación y anotación del corpus.
  - Ejercicio práctico 1: validación métrica.
- Segunda parte:
  - Modelos de análisis métrico y semántico.
  - Ejercicio práctico 2: extracción de datos a gran escala y análisis.

Marco metodológico

- Aproximación tradicional:
  - Análisis en profundidad de autores canónicos.
- Aproximación computacional:
  - Análisis automático de TODOS los autores.
  - Método distante, a gran escala o macro-análisis.

Distan Reading (Moretti)

- Buscar lo común en la Hª de la Literatura, no lo excepcional.
- Analizar amplios periodos como un todo.
- Método cuantitativo:
  - Obtención de datos: objetividad.
    - Plantea preguntas.
  - Análisis de datos.

Moretti (2007) *La literatura vista desde lejos*.

Macroanálisis (Jocker)

- Análisis inmanentista mediante métodos computacionales:
  - *Topic Modeling* (*text mining*).
  - Temas recurrentes de la novela del s. XIX en inglés y su tratamiento según sexo del autor.
    - Corpus: 3346 novelas (inglesa, irlandesa y norteamericana).

Jockers y Mimno (2013) “Significant Themes...”.

Table 1
Twenty-five useful features in distinguishing between two classes.

	Label	Male-authors	Female-authors
1	Female fashion	−0.2015	0.2614
2	Flowers and natural beauty	−0.1698	0.2203
3	Tears and sorrow	−0.1619	0.2101
4	Drawing rooms	−0.16	0.2076
5	Drink as in liquor and beer and tobacco	0.1489	−0.1932
6	Governesses and education of children	−0.1469	0.1906
7	Nurses for children	−0.1467	0.1904
8	Pistols and other guns	0.144	−0.1869
9	Children girls	−0.1374	0.1783
10	Pity	−0.1341	0.174
11	Children	−0.1333	0.173
12	Facial features	−0.1324	0.1719
13	Affection	−0.132	0.1712
14	Health and illness	−0.1314	0.1705
15	Landlords	−0.1301	0.1688
16	Men with guns	0.1298	−0.1684
17	Moments of confusion in battle	0.1292	−0.1677
18	Grief and sorrow	−0.1269	0.1646
19	Happiness	−0.1253	0.1627
20	Afternoon and tea time	−0.1243	0.1613
21	Swords and weapons	0.1241	−0.161
22	Male clothing	0.1234	−0.1601
23	Tea and coffee	−0.1232	0.1599
24	Soldiers	0.121	−0.157
25	Dear girls children creatures	−0.1198	0.1554

Análisis a gran escala

- Cambiar las preguntas:
  - Búsqueda de lo general.
  - Rasgos comunes de todo el periodo.
- Complementario al método tradicional.
- Ejemplo:
  - Culturomics (Michel, et al., 2011) y Google n-grams: https://books.google.com/ngrams

Análisis a gran escala

- ¿Cuáles son los gustos métricos generales de la sonetística áurea? ¿Son constantes o varían a lo largo del periodo?
- ¿Hay patrones métricos característicos de algún autor, escuela, periodo…?, ¿contrastan unos con otros? ¿Hay diferencias rítmico-métricas entre el Renacimiento y el Barroco? ¿Hay gustos comunes?
- ¿Se utilizan patrones métrico similares para tratar temas similares? ¿Hay un ritmo específico para el poema amoroso? ¿Y para el panegírico? ¿Hay relación entre la métrica y la semántica?
- …

Método computacional

- Extracción de datos:
  - Objetividad.
  - Compilación y anotación de un corpus de referencia.
- Análisis de datos:
  - Análisis de frecuencias.
  - Técnicas de *Text Mining*.

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Compilación y anotación de un corpus de sonetos del Siglo de Oro

erc European Research Council

Corpus

- Amplia colección de textos digitales.
- Compilado en función de unos **criterios**.
- **Representativo** de un hecho, periodo, fenómeno, etc.
- **Anotado** para representar aspectos lingüístico-literarios profundos.

Sinclair (2004) “Developing Linguistic Corpora...”

Corpus sonetos SdO

- Criterios de compilación:
  - Sonetos en castellano.
  - Siglos XVI y XVII.
- Criterios de anotación.
  - Metadatos.
  - Anotación estructural.
  - Anotación métrica.

Navarro Colorado et al (2016) “Metrical annotation...”

Compilación

- Textos extraídos de la Biblioteca Virtual Migue de Cervantes.
  - Ortografía y tipografía moderna.
- Autores con al menos 10 sonetos digitalizados.
- Corpus abierto.

Compilación

- Corpus abierto.
- Actualmente:
  - 52 poetas
  - 5077 sonetos
  - 71136 verso

https://github.com/bncolorado/CorpusSonetosSigloDeOro

Anotación

- Información a representar:
  - Metadatos: autor, título, fuente bibliográfica, etc.
  - Estructura: estrofas, versos.
  - Métrica: patrón métrico.

Formalismo

- XML - TEI.
- Encabezado TEI estándar:
  - Descripción del fichero.
    - Título y responsable, publicación, descripción de la edición fuente.
  - Codificación información métrica:
    - Definición formal de patrón métrico (expresión regular), anotación automática o manual.

https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/GarcilasoDeLaVega/GarcilasoDeLaVega_01.xml

Formalismo

- Cuerpo – estructura del poema:
  - Título del poema
  - Estrofa (lg): cuarteto, terceto, estrambote.
  - Versos (l, @ID).
  - Información métrica: patrón métrico.

https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/GarcilasoDeLaVega/GarcilasoDeLaVega_01.xml

Patrón métrico

- **Objetividad.**
- Secuencia de sílabas tónicas (+) y átonas (-) delimitadas por una pausa versal.
- Base del ritmo.
  - `<l n="1" met="---+---+-+-">`
  - `Cuando me paro a contemplar mi estado,`
  - `</l>`

https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/GarcilasoDeLaVega/GarcilasoDeLaVega_01.xml

Proceso de anotación

- **Consistencia, corrección y rapidez.**
- Consistencia y corrección:
  - Guía de anotación.
- Rapidez:
  - Proceso semi-automático.

Anotación Semiautomática

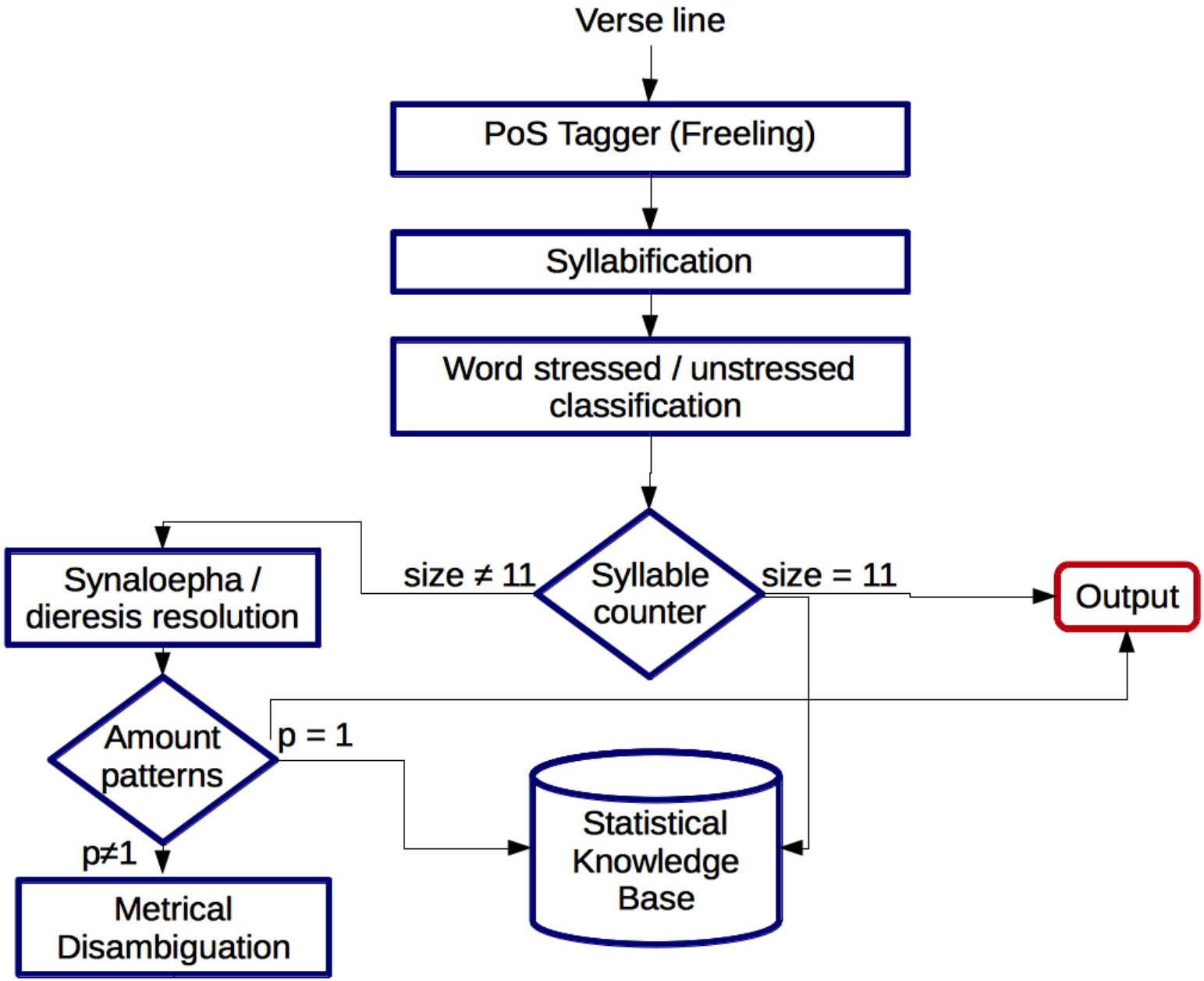
Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE iNNOVACIÓN EN HUMANIDADES DIGITALES

- Fase 1: Anotación automática.
- Fase 2: revisión manual.
  - Ciega y en paralelo.
  - Guía de anotación.
- Fase 3: revisión manual versos erróneos.

erc European Research Council

Fase automática

- Metadatos y estructura:
  - Python, XML y expresiones regulares.
- Métrica:
  - Sistema de escansión.


Verse line
PoS Tagger (Freeling)
Syllabification
Word stressed / unstressed classification
Syllable counter
size ≠ 11
size = 11
Output
Synaloepha / dieresis resolution
Amount patterns
p = 1
p≠1
Metrical Disambiguation
Statistical Knowledge Base

Ambigüedad métrica

- “cuando el padre Hebrero nos enseña”
  - (DiegoHurtadoDeMendoza_54.xml)
  - `<met =--+--+---+->`
  - `<met =---+-+---+->`

- Selecciona el más frecuente.

Fase manual

- Guía de anotación:
  - Qué hacer en casos dudosos.
- Entrenamiento: 500 versos.
- Validación 100 sonetos (1400 versos):
  - Dos anotadores más árbitro.

Evaluación

- Consistencia validación manual:
  - *Inter-annotators agreement*: 96%
- Precisión sistema automático:
  - 92%

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE iNNOVACIÓN EN HUMANIDADES DIGITALES
erc European Research Council

Fase final

- Validación versos ambiguos.
- Se evita la validación manual de todos los versos, pero se asegura la corrección y consistencia de la anotación final.

Conclusiones

- Corpus sonetos Siglo de Oro:
  - **Amplio** (más de 70 versos)
  - **Representativo**: cubre (casi) todos los autores del periodo.
  - Anotación métrica **objetiva** y **consistente**.
  - Creación (relativamente) rápida.
- Datos objetivos del periodo y fenómeno a estudiar.

Práctica 1

- Validación manual anotación métrica.
- Proceso:
  - Revisar guía de anotación (en materiales)
  - Validar hasta 5 sonetos (los que dé tiempo)
    - Incluid nombre del revisor en metadatos.
    - Documentad los cambios.
  - Cálculo de consistencia de la anotación.

Guía de anotación métrica (Quilis 84)

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

- Palabras tónicas:
  - Sustantivos, adjetivos, verbos, pronombres personales tónicos, indefinidos, demostrativos, posesivos, adverbios, formas interrogativas, numerales cardinales y ordinales...
- Palabras átonas:
  - Artículo determinado, preposición, conjunciones, fórmulas de tratamiento, pronombres personales átonos, determinantes posesivos, ...

erc European Research Council

Guía de anotación métrica

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE iNNOVACIÓN EN HUMANIDADES DIGITALES

- En caso de duda:
  - Preferencia por lectura átona
    - Acento secundario.
  - Buscar el patrón resultante más común, el menos forzado.
    - Evitar acentos antirrítmicos.

erc European Research Council

Ejemplo

- ...

¿Dudas?

- …

Universitat d'Alacant
Universidad de Alicante
Departament de Llenguatges i Sistemes Informàtics
Departamento de Lenguajes y Sistemas Informáticos
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

erc European Research Council

Bibliografía

- Garcíia Berrio, A. (2000) "Retórica figural. Esquemas argumentativos en los sonetos de Garcilaso". *Edad de Oro*, 19.
- Jockers, M. (2013), *Macroanalysis. Digital Media and Literary History*. University of Illinois Press, Illinois.
- Jockers, M. y Mimno, D. (2013) "Significant Themes in 19th-Century Literature", *Poetics*, 41.
- Michel, et al. (2011) "Quantitative Analysis of Culture Using Millions of Digitized Books" *Science* 331, 176.
- Moretti, F. (2007) *La literatura vista desde lejos*. Marbot ediciones.
- Moretti, F. (2013), *Distant reading*. Verso.
- Navarro Colorado, B. (2015) "A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects", *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, pages 105–113, Denver (Colorado), June 4, 2015.

Bibliografía (2)

- Navarro Colorado, B.; María Ribes-Lafoz, M. y Sánchez, N. (2016) "Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož (Slovenia).
- Padró, L y Stanilovsky, E. (2012) "FreeLing 3.0: Towards Wider Multilinguality". In *Proceedings of the International Conference Language Resources and Evaluation Conference* (LREC 2012), Istanbul.
- Quilis, A. (1984) *Métrica española*. Ariel, Barcelona.
- Rivers. E. (1993) *El soneto español en el siglo de oro*. Akal, Madrid.
- Sinclair, John (2004) "Developing Linguistic Corpora: a Guide to Good Practice" AHDS, http://users.ox.ac.uk/~martinw/dlc/chapter1.htm

Gracias

Contacto:

Borja Navarro Colorado
Dto. Lenguajes y Sistemas Informáticos
Universidad de Alicante
borja@dlsi.ua.es
@bncolorado