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

B. Navarro-Colorado[1,3], M. Ribes[2,3], S. Trigueros[3] y N. Sánchez[3]

[1]Dto. de Lenguajes y Sistemas Informáticos,
[2]Centro Superior de Idiomas
[3]Universidad de Alicante
borja@dlsi.ua.es, chitty@csidiomas.ua.es
{saraj.trigueros,noeliasanchezlopez}@gmail.com

Humanidades Digitales Hispánicas
Madrid, 5-7/X/2015

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

CENTRO SUPERIOR DE IDIOMAS

Índice

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

Índice

Objetivo del proyecto

- Aplicación de técnicas computacionales al análisis de la lírica castellana de los Siglos de Oro (ss. XVI-XVII).
  - Nuevo enfoque: *Distant Reading*
  - Nuevos métodos: Lingüística Computacional - *Text Mining*

Objetivo específico

Creación de un amplio corpus representativo de todo el periodo.

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

Objetivo del proyecto

- Aplicación de técnicas computacionales al análisis de la lírica castellana de los Siglos de Oro (ss. XVI-XVII).
  - Nuevo enfoque: *Distant Reading*
  - Nuevos métodos: Lingüística Computacional - *Text Mining*

Objetivo específico

Creación de un amplio corpus representativo de todo el periodo.

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
- Análisis de la “periferia contextual” [García Berrio, 2000].

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
  - Análisis de amplios periodos literarios como un todo.
  - Complementario a métodos tradicionales (*closed reading*)
  - Buscar lo general del hecho literario, no lo particular.
  - Métodos cuantitativos.
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
- Análisis de la “periferia contextual” [García Berrio, 2000].

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
  - Análisis de amplios periodos literarios como un todo.
  - Complementario a métodos tradicionales (*closed reading*)
  - Buscar lo general del hecho literario, no lo particular.
  - Métodos cuantitativos.
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
- Análisis de la “periferia contextual” [García Berrio, 2000].

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
  - Análisis de amplios periodos literarios como un todo.
  - Complementario a métodos tradicionales (*closed reading*)
  - Buscar lo general del hecho literario, no lo particular.
  - Métodos cuantitativos.
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
- Análisis de la “periferia contextual” [García Berrio, 2000].

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
  - Análisis de amplios periodos literarios como un todo.
  - Complementario a métodos tradicionales (*closed reading*)
  - Buscar lo general del hecho literario, no lo particular.
  - Métodos cuantitativos.
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
- Análisis de la “periferia contextual” [García Berrio, 2000].

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
  - Análisis computacional de amplios corpus de texto literario.
  - Aplicación de técnicas de Lingüística Computacional y Minería de Textos (*cluster*, semántica vectorial, *LDA Topic Modeling*, etc.).
- Análisis de la “periferia contextual” [García Berrio, 2000].

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
  - Análisis computacional de amplios corpus de texto literario.
  - Aplicación de técnicas de Lingüística Computacional y Minería de Textos (*cluster*, semántica vectorial, *LDA Topic Modeling*, etc.).
- Análisis de la “periferia contextual” [García Berrio, 2000].

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
- Análisis de la "periferia contextual" [García Berrio, 2000].
  - Periferia contextual de la cultura literaria de un autor.
  - Parámetros de interpretación global sobre el comportamiento individual.
  - Tratamiento informático masivo.

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
- Análisis de la “periferia contextual” [García Berrio, 2000].
  - Periferia contextual de la cultura literaria de un autor.
  - Parámetros de interpretación global sobre el comportamiento individual.
  - Tratamiento informático masivo.

Nuevo enfoque

- *Distant Reading* [Moretti, 2007, Moretti, 2013].
- *Macroanalysis* [Jockers, 2013, Jockers, 2014].
- Análisis de la "periferia contextual" [García Berrio, 2000].
  - Periferia contextual de la cultura literaria de un autor.
  - Parámetros de interpretación global sobre el comportamiento individual.
  - Tratamiento informático masivo.

Objetivo específico.

- Necesidad de amplios corpus de texto literario representativos del periodo.
- Creación de un corpus de sonetos del Siglo de Oro.
  - Análisis de aspectos métrico: corpus anotado.
  - Análisis de aspectos semánticos: corpus sin anotar.
- Compilación y anotación métrica del corpus de sonetos del Siglo de Oro.

Objetivo específico.

- Necesidad de amplios corpus de texto literario representativos del periodo.
- Creación de un corpus de sonetos del Siglo de Oro.
  - Análisis de aspectos métrico: corpus anotado.
  - Análisis de aspectos semánticos: corpus sin anotar.
- Compilación y anotación métrica del corpus de sonetos del Siglo de Oro.

Índice

Compilación: selección de textos.

- Corpus representativo de TODA la creación sonetística áurea.
  - No sólo autores canónicos.
  - Al menos **10 sonetos** digitalizados y disponibles.
  - Aprox. desde Garcilaso de la Vega (m. 1536) hasta Sor Juana Inés de la Cruz (m. 1695).
- Textos extraídos de la Biblioteca Virtual Miguel de Cervantes [García-González, 2007].
  `http://www.cervantesvirtual.com/`

Compilación: selección de textos.

- Corpus representativo de TODA la creación sonetística áurea.
  - No sólo autores canónicos.
  - Al menos **10 sonetos** digitalizados y disponibles.
  - Aprox. desde Garcilaso de la Vega (m. 1536) hasta Sor Juana Inés de la Cruz (m. 1695).
- Textos extraídos de la Biblioteca Virtual Miguel de Cervantes [García-González, 2007].
  `http://www.cervantesvirtual.com/`

Compilación: edición de los textos.

- No se pretende hacer una edición crítica de los sonetos.
- Detección de errores:
  - Cotejo con *editio princeps* (en BDH) y ediciones modernas.
  - Marcado de variantes relevantes.

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

Anotación

XML-TEI

- Encabezado: metadatos.
- Cuerpo:
  - Información estructural.
  - Información métrica.

Anotación

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.

Ejemplo

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title> Spanish Metrical Patterns Bank: Golden Age Sonnets.</title>
                <principal>Borja Navarro Colorado</principal>
                <respStmt>
                    <name>María Ribes Lafoz</name>
                    <name>Noelia Sánchez López</name>
                    <name>Borja Navarro Colorado</name>
                    <resp>Metrical patterns annotation</resp>
                </respStmt>
            </titleStmt>
            <publicationStmt>
                <publisher>Natural Language Processing Group. Department of Software and
Computing Systems. University of Alicante (Spain)</publisher>
            </publicationStmt>
            <sourceDesc>
                <bibl><title>Sonetos</title> de <author>Garcilaso de La Vega</author>.
<publisher>Biblioteca Virtual Miguel de Cervantes</publisher>, edición de <editor
role="editor">Ramón García González</editor>.</bibl>
            </sourceDesc>
        </fileDesc>
        <encodingDesc>
            <metDecl xml:id="bncolorado" type="met" pattern="((\+|\-)+)*">
                <metSym value="+">stressed syllable</metSym>
                <metSym value="-">unstressed syllable</metSym>
            </metDecl>
            <metDecl>
                <p>All metrical patterns have been manually checked.</p>
            </metDecl>
        </encodingDesc>
    </teiHeader>
    <text>
```

Anotación

Cuerpo:

- Información estructural.
  - Título.
  - Estrofa (*lg*): cuarteto, terceto (estrambote).
  - Versos (*l*, @*ID*).
- Información métrica.

Representación de la información métrica

- Sólo información objetiva.
  - No aspectos que pueden variar en ritmo o declamación: acentos secundarios, pausas potenciales, etc.
- Unidad métrica:
  - Estudios de métrica [Varela Merino et al., 2005]: **grupo métrico**.
    - Problema: carácter subjetivo de la pausa potencial.
  - Corpus: **verso**.
    - Delimitado por pausa “objetiva” (fin de verso).

Representación de la información métrica

- Sólo información objetiva.
  - No aspectos que pueden variar en ritmo o declamación: acentos secundarios, pausas potenciales, etc.
- Unidad métrica:
  - Estudios de métrica [Varela Merino et al., 2005]: **grupo métrico**.
    - Problema: carácter subjetivo de la pausa potencial.
  - Corpus: **verso**.
    - Delimitado por pausa “objetiva” (fin de verso).

Representación de la información métrica

- Sólo información objetiva.
  - No aspectos que pueden variar en ritmo o declamación: acentos secundarios, pausas potenciales, etc.
- Unidad métrica:
  - Estudios de métrica [Varela Merino et al., 2005]: **grupo métrico**.
    - Problema: carácter subjetivo de la pausa potencial.
  - Corpus: **verso**.
    - Delimitado por pausa “objetiva” (fin de verso).

Representación de la información métrica

- Patrón métrico: secuencia de sílabas tónicas (+) y átonas (−) delimitadas por una pausa versal.
  - Endecasílabos
- Base métrica objetiva sobre la que se realiza el ritmo y la declamación.

```
<l n="1" met="---+---+-+-">
   Cuando me paro a contemplar mi estado,
</l>
```

Ejemplo

```
<text>
    <body>
        <head>
            <title>-I-</title>
        </head>
        <lg type="cuarteto">
            <l n="1" met="---+---+-+-">Cuando me paro a contemplar mi estado,</l>
            <l n="2" met="-+-+---+-+-">y a ver los pasos por do me ha traído,</l>
            <l n="3" met="+--+--+--+-">hallo, según por do anduve perdido,</l>
            <l n="4" met="--++-+-+-+-">que a mayor mal pudiera haber llegado;</l>
        </lg>
        <lg type="cuarteto">
            <l n="5" met="-----+-+-+-">mas cuando del camino estó olvidado,</l>
            <l n="6" met="-+-+++-+-+-">a tanto mal no sé por do he venido;</l>
            <l n="7" met="+--+-+++-+-">sé que me acabo, y más he yo sentido</l>
            <l n="8" met="+--+-+---+-">ver acabar conmigo mi cuidado.</l>
        </lg>
        <lg type="terceto">
            <l n="9" met="+--+---+-+-">Yo acabaré, que me entregué sin arte</l>
            <l n="10" met="---+-+---+-">a quien sabrá perderme y acabarme</l>
            <l n="11" met="+--+---+-+-">si ella quisiere, y aun sabrá querello;</l>
        </lg>
        <lg type="terceto">
            <l n="12" met="-----++--+-">que, pues, mi voluntad puede matarme,</l>
            <l n="13" met="-+--++---+-">la suya, que no es tanto de mi parte,</l>
            <l n="14" met="-+-+-+---+-">pudiendo, ¿qué hará sino hacello?</l>
        </lg>
    </body>
</text>
</TEI>
```

Datos del corpus

- 52 Autores.
- 5078 sonetos: más de 71100 versos / patrones métricos.

Índice

Dilema en el desarrollo de recursos.

- Anotación automática: mucha cobertura, poca precisión.
- Anotación manual: buena precisión (calidad), poca cobertura (tamaño).

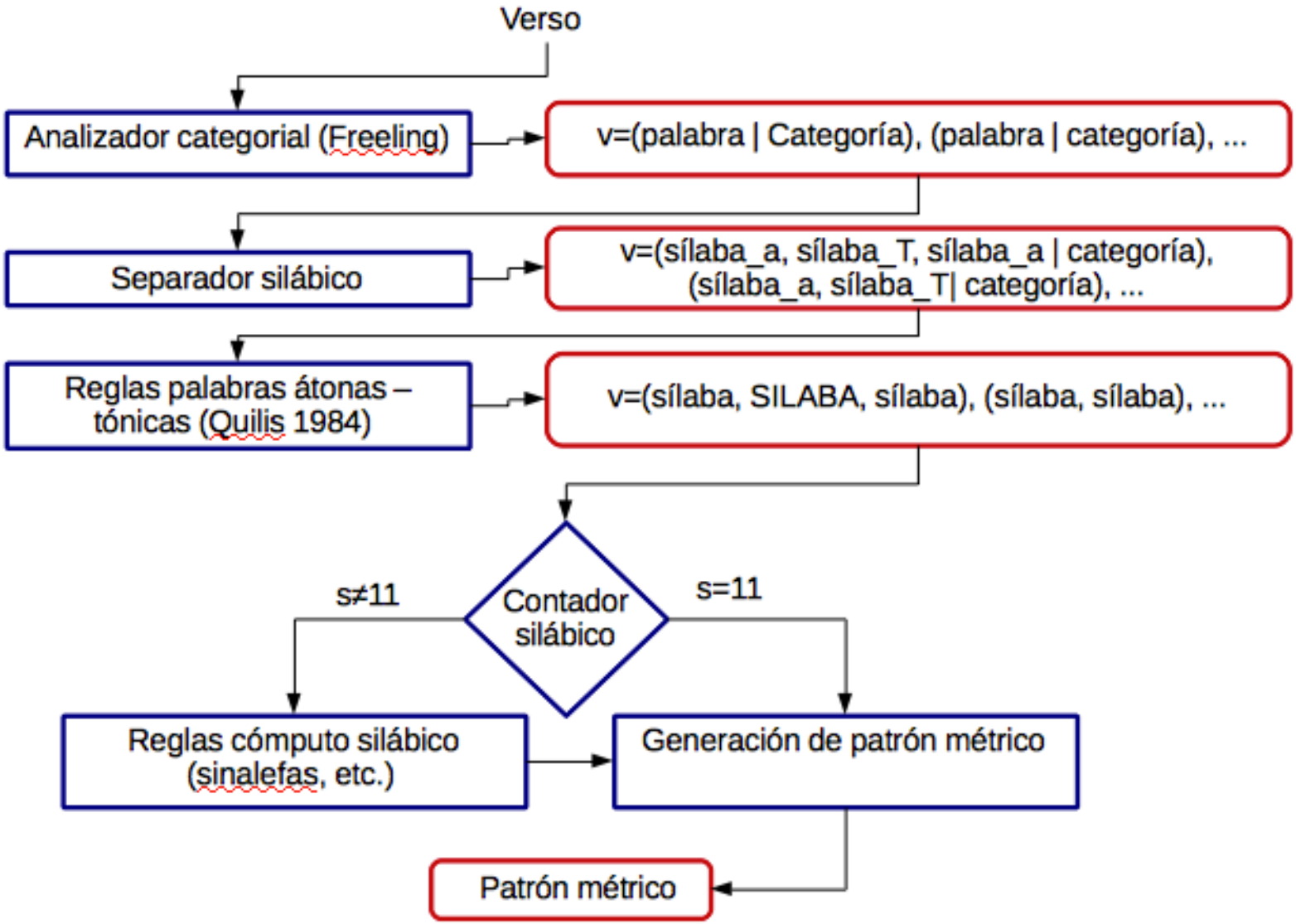
Proceso de anotación

Semiautomático:

- Anotación automática.
  - Encabezado TEI y anotación estructural.
  - Sistema de escansión específico.
- Validación manual.

Sistema de escansión métrica

- Basado en reglas.



Validación manual

Tres anotadores. 1 % del corpus validado.
Fases:

1. Entrenamiento. Guía de anotación.
2. Evaluación: acuerdo entre anotadores. (*)
3. Validación de versos potencialmente ambiguos.
4. Iteración.
   - Módulo de aprendizaje automático.
   - Ante métrica ambigua, selecciona el patrón más frecuente.

Principales problemas de anotación

- Errores del analizador categorial.
- Ambigüedad de las reglas de palabras acentuadas.
- Ambigüedad cómputo silábico.

Principales problemas de anotación

- Errores del analizador categorial.
  - Freeling. [Padró and Stanilovsky, 2012]
  - Ejemplo:

y	a	ver	los	pasos	por	do	me	ha	traído
y	*a*	*ver*	*el*	*paso*	*por*	*do*	*me*	*haber*	*traer*
CC	SPS00	VMN0000	DA0MP0	NCMP000	SPS00	NCMS000	PP1CS000	VAIP3S0	VMP00SM

Principales problemas de anotación

- Ambigüedad de las reglas de palabras acentuadas.
  - “locución breve de carácter vocativo y en expresiones cortas de cariño o reproche” son átonas [Navarro Tomás, 1974] (“Buen hombre”).
    - ¿Cuánto es «corto»?: máximo dos sílabas.
      a razonar con vos, ¡oh dulce amigo!, *met* $= - - - + - + - - - + -$
      (GarcilasDeLaVega_19.xml)

Principales problemas de anotación

- Ambigüedad de las reglas de palabras acentuadas.
  - “Un”
    - [Navarro Tomás, 1974]. Artículo indefinido: tónico.
    - [Spang, 1983]. Artículo indefinido: átono.
    - No queda claro cuándo es átono y cuándo tónico. Algunos versos parece que prefieren la lectura átona para evitar acentos anti-rítmicos.
    - “dejad un rato la labor, alzando” (GarcilasoDeLaVega_11.xml)
      $< met = - + + + - - - + - + - >$
      $< met = - + - + - - - + - + - >$

Principales problemas de anotación

Ambigüedad cómputo silábico:

- Versos que permiten más sinalefas/sinéresis de las necesarias.
  - “cuan**do el** pa**dre He**brero nos enseña” (DiegoHurtadoDeMendoza_54.xml)
    $< met = - - + - - + - - - + - >$
    $< met = - - - + - + - - - + - >$
- Seleciona el más frecuente.

Índice

Patrones métricos más frecuentes

Patrón métrico	Nombre	Apariciones
- + - - - + - - - + -	Heroico	**6457**
- + - + - - - + - + -	Sáfico	**6161**
- - + - - + - - - + -	Melódico	**5982**
- + - + - + - - - + -	Heroico	5015
- - - + - + - - - + -	Sáfico	3947
- + - - - + - + - + -	Heroico	3549
- + - + - + - + - + -	Heroico	3310
+ - - + - - - + - + -	Sáfico	3164
+ - - + - + - - - + -	Sáfico	3150
- - - + - - - + - + -	Sáfico	3105
- - + - - + - + - + -	Melódico	2940

Análisis semánticos (provisional)

LDA Topic Modeling [Blei et al., 2003] sobre el texto.

- Tipos de *topics*:
  - Relacionados con temas clásicos (amor no correspondido, funeral, decadencia del imperio).
  - Rimas.
  - Relaciones figurativas: “río” - “cristal”.
  - Ruido.

Análisis semánticos (provisional)

LDA Topic Modeling sobre el texto.

- *Cluster* de autores según *topics* comunes ($t = 20$, $n = 10$).
  - Íñigo López de Mendoza.
  - Hernando de Acuña, Juan de Timoneda, Juan Boscán, Garcilaso de la Vega, Gutierre de Cetina, Diego Hurtado de Mendoza (Renacimiento 1).
  - Miguel de Cervantes, Fray Luis de León, Francisco de Figueroa, Francisco de la Torre, Francisco de Aldana, Juan de Almeida (Renacimiento 2).
  - Fernando de Herrera.
  - Resto de grupos: poetas barrocos.

Análisis métricos (provisional)

LDA Topic Modeling sobre patrones métricos.

- 10 *topics* y 4 grupos.
  - Íñigo López de Mendoza.
  - Poetas renacentistas (excepto Mira de Amescúa y Luis Carrillo y Sotomayor).
  - Poetas barrocos (inc. Lope de Vega y Cervantes).
  - Poetas barrocos (inc. Góngora y Quevedo).

Índice

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

Conclusiones

- Macroanálisis sobre lírica de los Siglos de Oro.
- Necesidad de amplios corpus anotados: procesos semiautomáticos.
- Representación formal de la información métrica y sistema automático de escasión.
- Corpus disponible en:

`http://www.dlsi.ua.es/~borja/mpb/`

Trabajo futuro

- Corto plazo
  - Finalizar la validación manual de versos ambiguos (resto).
  - Mejorar sistema de escansión con métodos estadísticos.
- Medio/largo plazo
  - Ampliar el corpus con más autores.
  - Incorporar ediciones críticas digitales (libres y disponibles).
  - Ampliar y adaptar a otros tipos de poemas (romances).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent Dirichlet Allocation.
Journal of Machine Learning Research, 3:993–1022.

García Berrio, A. (2000).
Retórica figural. Esquemas argumentativos en los sonetos de Garcilaso.
Edad de Oro, (19).

García-González, R., editor (2007).
Biblioteca del soneto.
Biblioteca Virtual Miguel de Cervantes, Alicante.

Jockers, M. L. (2013).
Macroanalysis. Digital Media and Literary History.
University of Illinois Press, Illinois.

Jockers, M. L. (2014).
Text Analysis with R for Students of Literature.
Springer.

Moretti, F. (2007).
Graphs, Maps, Trees: Abstract Models for a Literary History.
Verso.

Moretti, F. (2013).
Distant reading.
Verso.

Navarro Tomás, T. (1974).
Manual de pronunciación española.
Consejo Superior de Investigaciones Científicas, Madrid.

Padró, L. and Stanilovsky, E. (2012).
FreeLing 3.0: Towards Wider Multilinguality.
In *Language Resources and Evaluation Conference (LREC 2012)*, Istanbul.

Spang, K. (1983).
Ritmo y versificación. Teoría y práctica del análisis métrico y rítmico.
Universidad de Murcia, Murcia.

Varela Merino, E., Moíno Sánchez, P., and Jauralde Pou, P. (2005).
Manual de métrica española.
Castalia, Madrid.