Edición crítica digital y anotación de corpus para estudios literarios computacionales

Borja Navarro Colorado

Filología Digital e Inteligencia Artificial
Burgos, 5 de septiembre de 2023

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

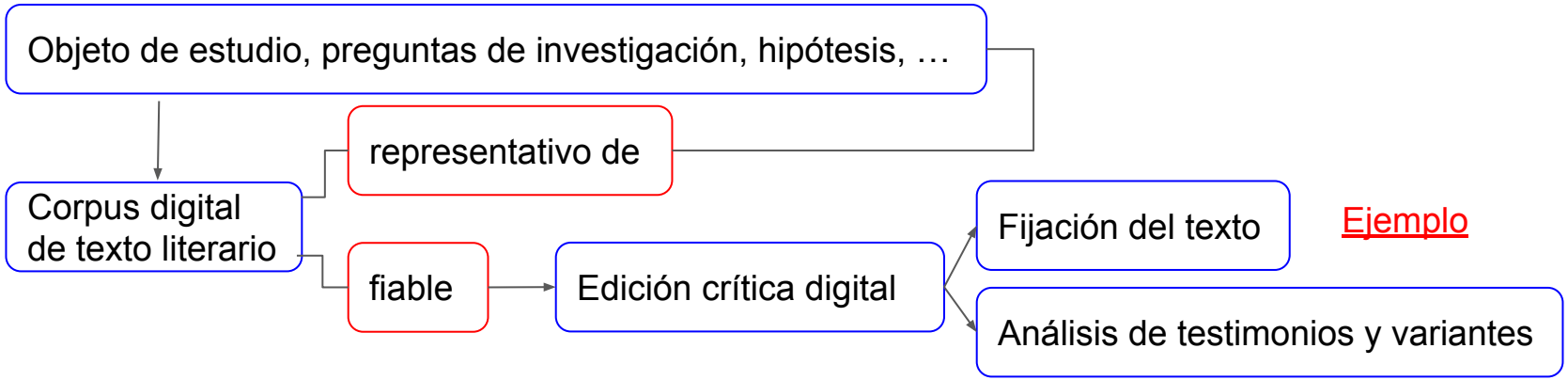
GPLSI

Financiado Proyecto "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00), Ministerio de Ciencia e Innovación MCIN/ AEI/10.13039/501100011033 y proyecto NL4DISMIS: Tecnologías del Lenguaje Natural para lidiar con la desinformación (CIPROM/2021/021), Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport)

No preguntes
qué puede hacer la tecnología por ti,
sino qué puedes hacer tú
por la tecnología

Adaptar
la herramienta computacional
al estudio literario,
y no al revés.

Estudios literarios computacionales



Ejemplos de edición crítica digital

```
<listWit>
    <witness xml:id="A">
        <msDesc type="manuscript">
            <msIdentifier>
                <country>España</country>
                <settlement>Madrid</settlement>
                <repository>Biblioteca Nacional de España</repository>
                <collection>Fondo antiguo</collection>
                <idno type="cataloguenumber">Mss/4075</idno>
                <msName>Varias poesías de Luis de Góngora</msName>
                <msName>Manuscrito Alba</msName>
            </msIdentifier>
            <msContents>
                <msItem>
                    <locus from="235r" to="275v">235r-275v</locus>
                    <title level="m">Soledades</title>
                </msItem>
            </msContents>
        </msDesc>
    </witness>
    <witness xml:id="Br">
        <msDesc type="manuscript">
```

```
<l xml:id="v-0007" exclude="#Rm #Pr">bates los montes que, de nieve armados,</l>
<l xml:id="v-0008" exclude="#Rm #Pr">gigantes de cristal los teme el cielo,</l>
<l xml:id="v-0009" exclude="#Rm #Pr">donde el cuerno, <app>
        <lem>del eco repetido,</lem>
        <rdg type="error" wit="#Rl">de eco repetido,</rdg>
    </app>
</l>
<l n="10" xml:id="v-0010" exclude="#Rm #Pr">fieras te expone que, al <app>
        <lem>teñido suelo,</lem>
        <rdg cert="low" type="error" wit="#Rl">temido suelo,</rdg>
    </app>
</l>
<l xml:id="v-0011" exclude="#Rm #Pr">
    <app>
        <lem>muertas,</lem>
        <rdg type="error" wit="#A">mu<add place="above">e</add>rtas,</rdg>
    </app> pidiendo términos deformes,</l>
<l xml:id="v-0012" exclude="#Rm #Pr">espumoso coral le dan al Tormes!:</l>
```

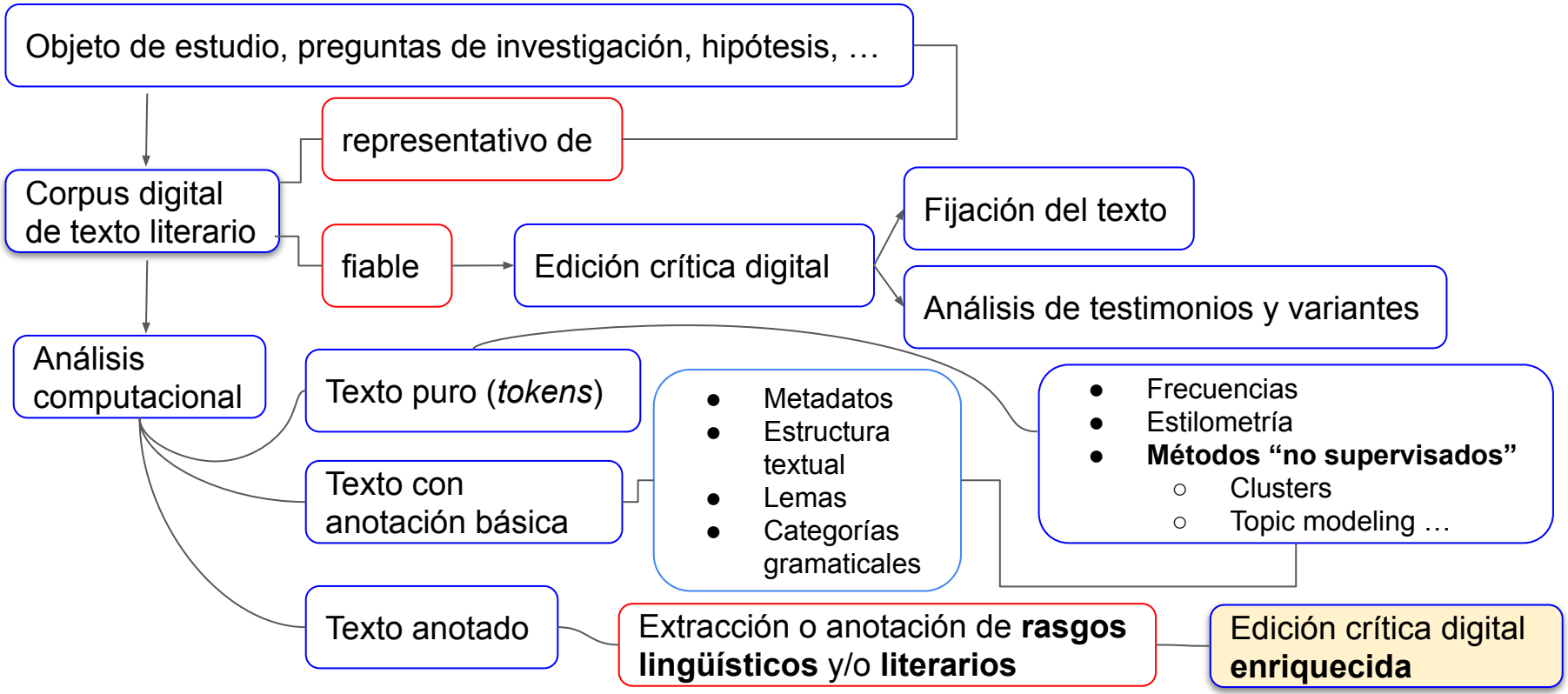
Góngora *Soledades* Edición crítica de A. Rojas Castro, en https://github.com/arojascastro/soledades

Ejemplos de edición crítica digital

[Colección digital Proteo]().

- Edición digital de las obras completas de Agustín Moreto
- Fijación del texto: contrastado y modernizado.
- Aparato crítico en PDFs.

Estudios literarios computacionales



Ejemplos de texto literario anotado

```
<lg type="cuarteto">
        <l n="1" met="-+-+-+---+-">Cerrar podrá mis ojos la postrera</l>
        <l n="2" met="+----+-+-+-">sombra que me llevare el blanco día,</l>
        <l n="3" met="--+--+++-+-">y podrá desatar esta alma mía</l>
        <l n="4" met="+--+-+---+-">hora a su afán ansioso lisonjera;</l>
</lg>
<lg type="cuarteto">
        <l n="5" met="-+++-+---+-">mas no, de esa otra parte, en la ribera,</l>
        <l n="6" met="--+--+---+-">dejará la memoria, en donde ardía:</l>
        <l n="7" met="-++--+-+-+-">nadar sabe mi llama el agua fría,</l>
        <l n="8" met="--+--+-+-+-">y perder el respeto a ley severa.</l>
</lg>
<lg type="terceto">
        <l n="9" met="+--+++-+++-">Alma a quien todo un dios prisión ha sido,</l>
        <l n="10" met="+--+-+-+++-">venas que humor a tanto fuego han dado,</l>
        <l n="11" met="-+-+-+-+-+-">medulas que han gloriosamente ardido,</l>
</lg>
<lg type="terceto">
        <l n="12" met="-+---++--+-">su cuerpo dejará, no su cuidado;</l>
        <l n="13" met="-+-+---+-+-">serán ceniza, mas tendrá sentido;</l>
        <l n="14" met="+--+-+---+-">polvo serán, mas polvo enamorado.</l>
</lg>
```

Navarro Colorado (ed) *Corpus de sonetos del Siglo de Oro*. https://github.com/bncolorado/CorpusSonetosSigloDeOro/tree/master
Vid. Navarro Colorado et al. 2016

Anotación automática del corpus

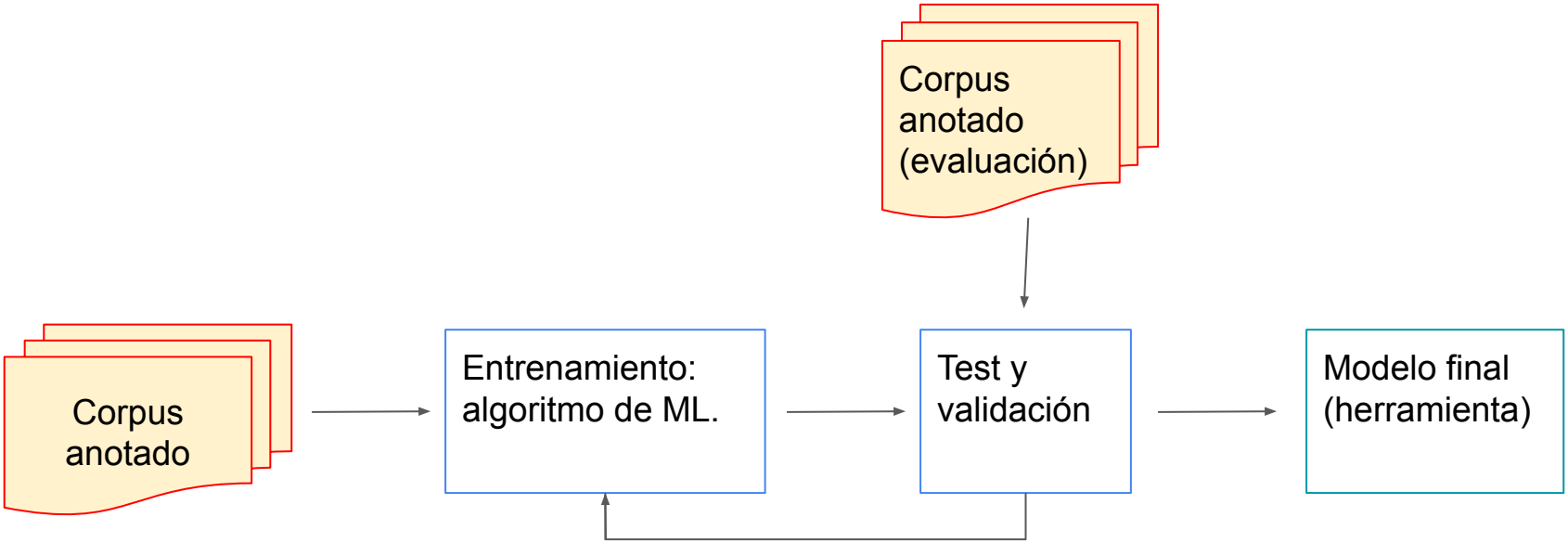
Herramientas de Procesamiento del Lenguaje Natural (PLN) o Inteligencia Artificial (IA)

- **¡Cuidado!** La mayoría de las herramientas de PLN/IA no están pensadas para estudios literarios.
- Problemas:
  1. Especificidad del texto literario:
     - Texto antiguo, histórico, no estándar.
     - Complejo: tensión idiomática, máxima capacidad expresiva (especificidad literaria).
  2. Análisis de aspectos propios de la literatura: métrica, personajes, secuencias de eventos narrativos, metáforas novedosas, etc. etc. etc.

Técnicas PLN/IA

- Reglas:
  - Técnica tradicional, aún útil en algunos casos. Ej. escansión.
- Aprendizaje automático (*machine learning*) clásico:
  - Métodos supervisados
  - Métodos no supervisados.
- Modelos neuronales (*deep learning*).

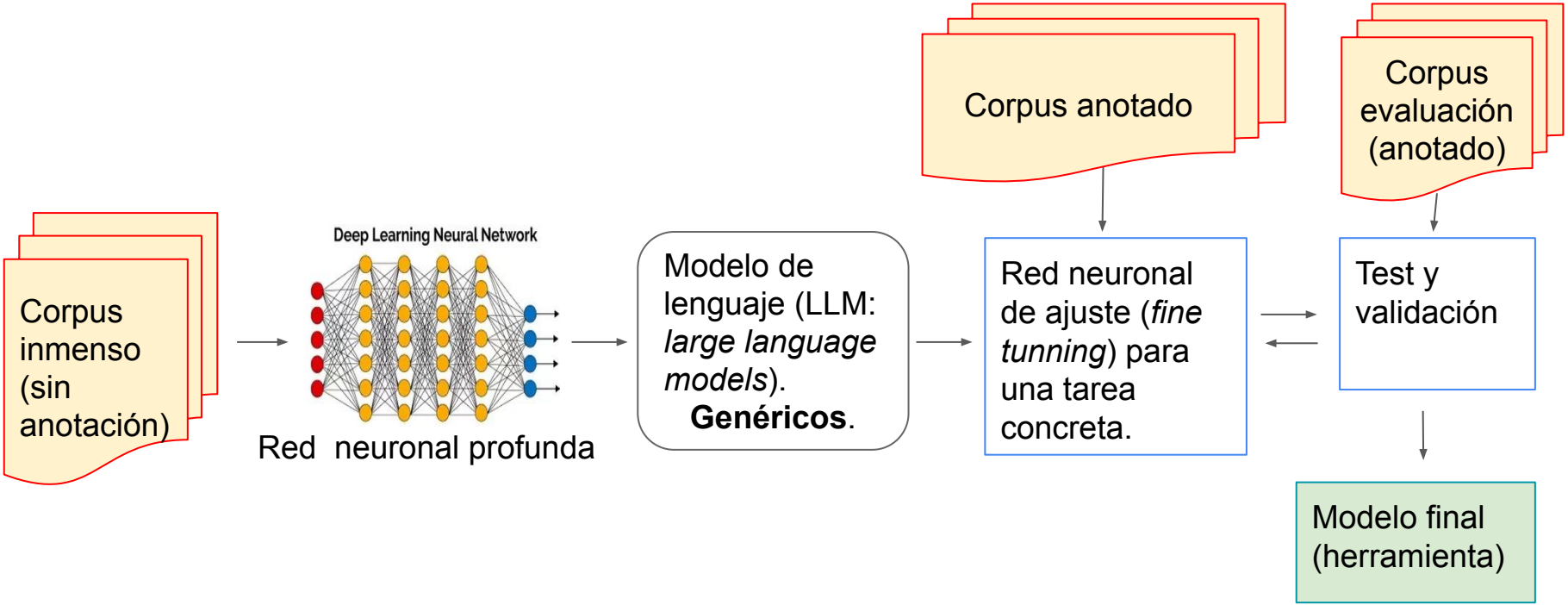
Aprendizaje automático supervisado



Técnicas PLN/IA

- Reglas:
  - Técnica tradicional, aún útil en algunos casos. Ej. escansión.
- Aprendizaje automático (*machine learning*) clásico:
  - Métodos supervisados
  - Métodos no supervisados.
- Modelos neuronales (*deep learning*).

Amplios modelos con *deep learning* (redes neuronales)

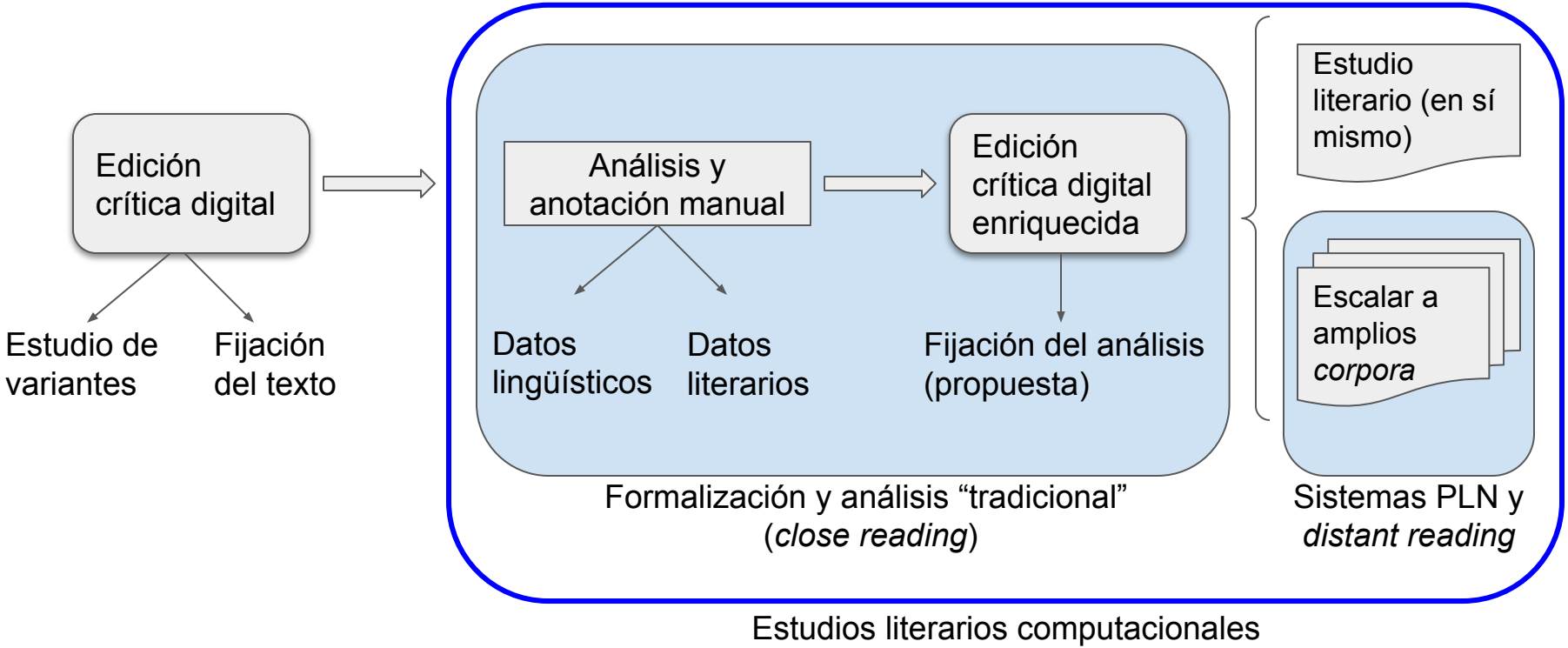


Técnicas de PLN/IA

Necesidad de corpus anotados o *data sets*:

- **Edición crítica digital enriquecida**.
- Aprendizaje para métodos supervisados, semi-supervisados o para *fine-tuning*.
- Evaluación y test para todos los casos.
- En estudios literarios computacionales como estudio en sí mismo.

ELC y Edición crítica enriquecida (recap)



Anotación de un corpus literario

- **¿Qué información anotar?**
- Formalización de la información.
- Métodos de anotación.

¿Qué información anotar?

Cualquier tipo de información que se quiera hacer explícita en el texto:

- Literaria: personajes, estilo indirecto libre, referencias mitológicas, métrica, metáforas, figuras retóricas, etc.
- Lingüística: categorías gramaticales, lemas, papeles semánticos, eventos, sentimientos, opiniones, ironía, etc.
- ...

¿Qué información anotar?

¿Existe una herramienta de anotación automática para mi tema?

¿Alguna herramienta que me pueda ayudar en mi tema?

- CLARIAH-ES: https://ixa2.si.ehu.eus/red-clariah-es/home
- CLARIN-EU: https://www.clarin.eu/
- DARIAH-EU: https://www.dariah.eu/

Desde el 1 de septiembre de 2023
formamos parte de la infraestructura europea
ERC CLARIN y DARIAH
https://www.clarin.eu/news/spain-joins-clarin-member

Anotación de un corpus literario

- ¿Qué información anotar?
- **Formalización de la información.**
- Métodos de anotación.
  - Automática (Procesamiento del Lenguaje Natural).
  - Manual.
  - Semi-automática: *active learning*.

Formalización de la información

Para que la máquina pueda “comprender” la información literaria y lingüística, debe ser formalizada:

1. Sistematización de la información: datos, cuantificación.
2. Representada explícitamente mediante un lenguaje formal.

Formalización de la información

Para que la máquina pueda “comprender” la información literaria y lingüística, debe ser formalizada:

1. Sistematización de la información: teoría de la literatura, estado de la cuestión…
    a. Temas e imágenes recurrentes soneto del Siglo de Oro: García Berrio 1978, Rivers 1993, Manero Sorolla 1990, Rosa Romojaro 2019, etc.
    b. Métrica del endecasílabo: Navarro Tomás (1956, 1995), Quilis (1993, 1984), Jauralde Pou (2020), etc.
2. Representada explícitamente mediante un lenguaje formal.

Lenguajes formales

Lenguajes de marcado: lenguaje formal para codificar un documento mediante **etiquetas**.

Tipos:

- Basado en SGML (*Standard Generalized Markup Language*)
  - Etiquetas <...>. Ej. casa
  - Lenguajes: HTML y **XML**.
- No basados en SGML.
  - LaTeX, MarkDown, mediaWiki, etc.

XML

Los corpus se suelen marcar con el lenguaje XML.

- Permite definir etiquetas propias (DTD o Schema).
- Etiquetas simples:
  - <title>La Celestina</title>
- Etiquetas complejas (atributo - valor)
  - <verso type=“endecasilabo”>Un soneto me manda hacer Violante</verso>

Ejemplos de texto literario anotado

```
<lg type="cuarteto">
        <l n="1" met="-+-+-+---+-">Cerrar podrá mis ojos la postrera</l>
        <l n="2" met="+----+-+-+-">sombra que me llevare el blanco día,</l>
        <l n="3" met="--+--+++-+-">y podrá desatar esta alma mía</l>
        <l n="4" met="+--+-+---+-">hora a su afán ansioso lisonjera;</l>
</lg>
<lg type="cuarteto">
        <l n="5" met="-+++-+---+-">mas no, de esa otra parte, en la ribera,</l>
        <l n="6" met="--+--+---+-">dejará la memoria, en donde ardía:</l>
        <l n="7" met="-++--+-+-+-">nadar sabe mi llama el agua fría,</l>
        <l n="8" met="--+--+-+-+-">y perder el respeto a ley severa.</l>
</lg>
<lg type="terceto">
        <l n="9" met="+--+++-+++-">Alma a quien todo un dios prisión ha sido,</l>
        <l n="10" met="+--+-+-+++-">venas que humor a tanto fuego han dado,</l>
        <l n="11" met="-+-+-+-+-+-">medulas que han gloriosamente ardido,</l>
</lg>
<lg type="terceto">
        <l n="12" met="-+---++--+-">su cuerpo dejará, no su cuidado;</l>
        <l n="13" met="-+-+---+-+-">serán ceniza, mas tendrá sentido;</l>
        <l n="14" met="+--+-+---+-">polvo serán, mas polvo enamorado.</l>
</lg>
```

Navarro Colorado (ed) *Corpus de sonetos del Siglo de Oro*. https://github.com/bncolorado/CorpusSonetosSigloDeOro/tree/master
Vid. Navarro Colorado et al. 2016

Estándar TEI

Text Encoding Initiative: https://tei-c.org/

Estructura general de un fichero TEI:

- Encabezado (<*teiHeader*>):
  - Metadatos como título, autor, datos bibliográficos, codificación, historial de revisiones, etc.
- Cuerpo (<*text*>):
  - Estructura de la obra: volúmenes, capítulos, párrafos, etc.
  - Citas, versos, salto de página, notas, cambio de idioma, énfasis, etc.

Ejemplo completo:

https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/Quevedo/Quevedo_142.xml

Estándar TEI

Más info:

- Guía oficial:
  - https://tei-c.org/Guidelines/P5/
- Recursos para aprender TEI en español:
  - https://tthub.io/aprende/
- Aprender mediante ejemplos:
  - http://www.teibyexample.org/

Problemas

COMPUTACIONALES

Edición crítica digital enriquecida (corpus literario anotado):

```
<lg type="cuarteto">
<l n="1" met="-+-+-+---+-">Cerrar podrá mis ojos la postrera</l>
 <l n="2" met="+----+-+-+-">sombra que me llevare el blanco día,</l>
<l n="3" met="--+--+++-+-">y podrá desatar esta alma mía</l>
<l n="4" met="+--+-+---+-">hora a su afán ansioso lisonjera;</l>
```

ESTUDIOS LITERARIOS

Fijación del análisis (interpretación)

↓

Permitir análisis alternativos

↓

Perspectivismo

Perspectivismo

- En anotación de corpus para PLN, la calidad de la anotación se mide por el acuerdo entre anotadores (Medida IAA: *inter-annotators agreement*).
  a. Diferentes medidas de cálculo (Kappa, …)
  b. A mayor acuerdo, más consistente la anotación y mayor calidad del corpus.
- Nuevo modelo: “perspectivismo”.
  a. El desacuerdo en la anotación no es un error, sino producto de diferentes interpretaciones: diferentes puntos de vista sobre los mismos fenómenos lingüísticos (o literarios).
  b. El desacuerdo es información muy valiosa porque:
     i. Indica los temas, fenómenos, aspectos a investigar (puntos calientes)
     ii. Un sistema automático debe aprender también que ese fenómeno no tiene una interpretación única.
  c. Manifiesto, bibliografía y corpus que siguen este modelo: http://pdai.info/

Problemas

COMPUTACIONALES

Compatibilidad con otros corpus y herramientas computacionales.

Anotación multinivel: representar diferentes tipos de análisis / anotación en el mismo corpus

Edición crítica digital enriquecida (corpus literario anotado):

```
<lg type="cuarteto">
<l n="1" met="-+-+-+---+-">Cerrar podrá mis ojos
la postrera</l>
 <l n="2" met="+----+-+-+-">sombra que me
llevare el blanco día,</l>
<l n="3" met="--+--+++-+-">y podrá desatar esta
alma mía</l>
<l n="4" met="+--+-+---+-">hora a su afán ansioso
lisonjera;</l>
```

ESTUDIOS LITERARIOS

Fijación del análisis (interpretación)

↓

Permitir análisis alternativos

↓

Perspectivismo

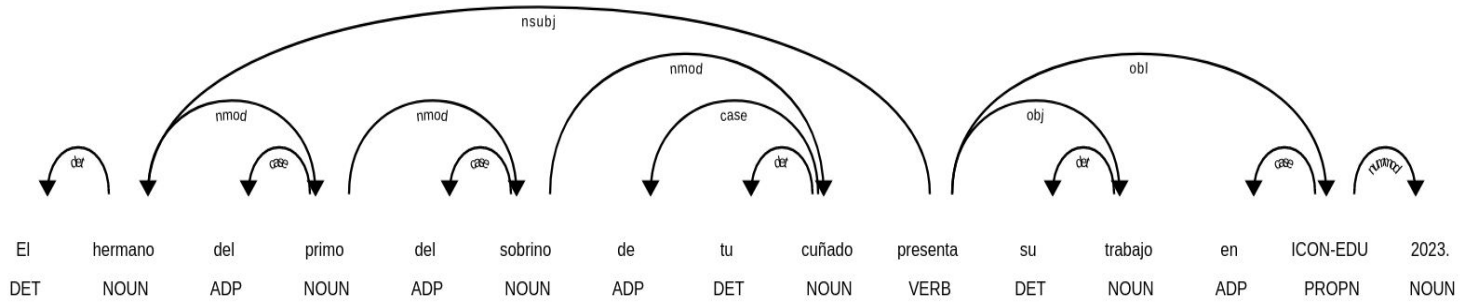
Compatibilidad con otros corpus y herramientas

Anotar con un lenguaje formal estandarizado: XML-TEI, pero:

- XML no es apropiado para algunos rasgos lingüísticos o literarios.
  - **Sintaxis**: formato CONLL (estándar *de facto*) o formato BIO.
- TEI no cubre todos los fenómenos lingüísticos o literarios.
  - Déficit para representación lingüística (de hecho no se usa en PLN).

Anotación información sintáctica - Formalismo CONLL

- En PLN, hoy el estándar es:
  - Análisis de dependencias
  - Modelo ***Universal Dependencies***. https://universaldependencies.org/
  - Formato CONLL
- Problema XML-TEI para representar grafos.



Anotación información sintáctica - Formalismo CONLL

```
1 Un    uno     DI0MS0  DI  - - - - 2 spec    - -
2 soneto   soneto   NCMS000 NC  - - - - 5 obj    - -
3 me    me      PP1CS00 PP  - - - - 4 obl     - -
4 manda mandar   VMIP3S0 VMI - - - - 0 sentence - -
5 hacer hacer   VMN0000 VMN - - - - 4 obj     - -
6 Violante violante NP00000 NP  - - - - 4 suj    - -
7 .     .       Fp      Fp  - - - - 4 f    - -
```

syntax_conll.txt

Anotación información sintáctica - Formalismo TEI

```
<graph type="directed" xml:id="RDG1">
<node n="1"><label>
     <w lemma="uno" pos="DI0MS0" msd="DI">Un</w>
</label></node>
<node n="2"><label>
     <w lemma="soneto" pos="NCMS000" msd="NC">soneto</w>
</label></node>
<node n="3"><label>
     <w lemma="me" pos="PP1CS00" msd="PP">me</w>
</label></node>
<node n="4"><label>
     <w lemma="mandar" pos="VMIP3S0" msd="VMI">manda</w>
</label></node>
<node n="5"><label>
     <w lemma="hacer" pos="VMN0000" msd="VMN">hacer</w>
</label></node>
<node n="6"><label>
     <w lemma="Violante" pos="NP00000" msd="NP">Violante</w>
</label></node>
```

```
<arc from="#1" to="#2">
     <label>spec</label>
</arc>
<arc from="#2" to="#5">
     <label>obj</label>
</arc>
<arc from="#3" to="#4">
     <label>obl</label>
</arc>
<arc from="#4" to="#0">
     <label>sentence</label
>
</arc>
<arc from="#5" to="#4">
     <label>obj</label>
</arc>
<arc from="#6" to="#4">
     <label>subj</label>
</arc></graph>
```

syntax_tei.xml

Anotación multinivel

Muchos fenómenos literarios requieren anotar el corpus con información diversa, a diferentes niveles de descripción lingüística y/o literaria.

- Diferentes unidades lingüísticas.
- **Problema cruce de ramas.**

Problema cruce de ramas

```
Así Fabio lloraba. Albania entonces
mirole, y quiso hablar, cerró los ojos,
y respondiole lo demás la muerte.
```

Lope de Vega “¿A dónde vas con alas tan ligeras,”
https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/LopeDeVega_1/LopeDeVega_125.xml

Problema cruce de ramas

```
<l>Así Fabio lloraba. Albania entonces</l>
<l>mirole, y quiso hablar, cerró los ojos,</l>
<l>y respondiole lo demás la muerte.</l>
```

Lope de Vega “¿A dónde vas con alas tan ligeras,”
https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/LopeDeVega_1/LopeDeVega_125.xml

Problema cruce de ramas

```
<l><s>Así Fabio lloraba.</s><s>Albania entonces</l>
<l>mirole, y quiso hablar, cerró los ojos,</l>
<l>y respondiole lo demás la muerte.</s></l>
```

Lope de Vega “¿A dónde vas con alas tan ligeras,”
https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/LopeDeVega_1/LopeDeVega_125.xml

Problema cruce de ramas

Soluciones TEI (1): anotar cada nivel en ficheros diferentes.

```
<l>Así Fabio lloraba. Albania entonces </l>
<l>mirole, y quiso hablar, cerró los ojos, </l>
<l>y respondiole lo demás la muerte. </l>
```

```
<s>Así Fabio lloraba.</s>
<s>Albania entonces mirole, y quiso hablar,
cerró los ojos,y respondiole lo demás la
muerte.</s>
```

Lope de Vega “¿A dónde vas con alas tan ligeras,”
https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/LopeDeVega_1/LopeDeVega_125.xml

Problema cruce de ramas

Soluciones TEI (2): Marcar solo el límite del segmento con una etiqueta simple <.../>

```
<l>Así Fabio lloraba.<s n=”1”/> Albania entonces</l>
<l>mirole, y quiso hablar, cerró los ojos, </l>
<l>y respondiole lo demás la muerte. <s n=”2”/></l>
```

Lope de Vega “¿A dónde vas con alas tan ligeras,”
https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/LopeDeVega_1/LopeDeVega_125.xml

Problema cruce de ramas

Soluciones TEI (3): Marcar solo el límite del segmento con una etiqueta simple <.../>

```
<l><s n="1">Así Fabio lloraba.</s><s n="2">Albania entonces</s></l>
<l><s n="2">mirole, y quiso hablar, cerró los ojos, </s></l>
<l><s n="2">y respondiole lo demás la muerte. </s></l>
```

Lope de Vega “¿A dónde vas con alas tan ligeras,”
https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/LopeDeVega_1/LopeDeVega_125.xml

Opciones TEI para anotación multinivel

Soluciones TEI (4): *stand-off Markup*:

- Separar el texto de la anotación en fichero diferentes.
- Indicar el *span* de texto de cada etiqueta.
- XML eficiente pero complejo y difícil de entender.

source.xml

```
Así Fabio lloraba. Albania entonces
mirole, y quiso hablar, cerró los ojos,
y respondiole lo demás la muerte.
```

metre.xml

```
<l><xi:include href="source.xml"
pointer="string-range(element(/1),0,35)"/></l>
<l><xi:include href="source.xml"
xpointer="string-range(element(/1),36,75)"/></l>
```

Lope de Vega “¿A dónde vas con alas tan ligeras,”
https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/LopeDeVega_1/LopeDeVega_125.xml

Anotación multinivel en CLS - Caso 1

Corpus General de Poesía Lírica del Siglo de Oro (Navarro-Colorado 2019)

- 23 poetas del Siglo de Oro (de Ercilla a Sta. Teresa)
- Diferentes tipos de estrofas: canciones, coplas, églogas, endechas, madrigales…
- Anotación TEI **métrica** y **categorial** **en un único fichero**.

```
<l met="-|-|-|-|5|-|7|-|" n="1">
    <!--A mis soledades voy,-->
    <w lemma="a" type="SP">a|</w>
    <w lemma="mi" type="DP1CPS">mis|</w>
    <w lemma="soledad" type="NCFP000">so|le|da|des|</w>
    <w lemma="ir" type="VMIP1S0">voy|</w>
</l>
```

https://github.com/bncolorado/CorpusGeneralPoesiaLiricaCastellanaDelSigloDeOro

Anotación multinivel en CLS - Caso 1

```
<l met="----+-+-" n="1">A mis soledades voy</l>
```

```
<l met="----+-+-" n="1">
    <w lemma="a" type="SP">a</w>
    <w lemma="mi" type="DP1CPS">mis</w>
    <w lemma="soledad" type="NCFP000">soledades</w>
    <w lemma="ir" type="VMIP1S0">voy</w>
</l>
```

```
<l met="-|-|-|-|5|-|7|-|" n="1">
    <!--A mis soledades voy,-->
    <w lemma="a" type="SP">a|</w>
    <w lemma="mi" type="DP1CPS">mis|</w>
    <w lemma="soledad" type="NCFP000">so|le|da|des|</w>
    <w lemma="ir" type="VMIP1S0">voy|</w>
</l>
```

https://github.com/bncolorado/CorpusGeneralPoesiaLiricaCastellanaDelSigloDeOro

Anotación multinivel en CLS - Caso 1

```
<l met="+----+-+-" n="2">prado de bienandanza que ni al yelo</l>
```

https://github.com/bncolorado/CorpusGeneralPoesiaLiricaCastellanaDelSigloDeOro/blob/master/FrayLuisDeLeon/frayLuisDeLeon_AlmaRegionLuciente.xml

Anotación multinivel en CLS - Caso 1

```
<l met="+----+---+-" n="2">
        <!--prado de bienandanza, que ni al yelo-->
        <w lemma="prado" type="NCMS000"> prado</w>
        <w lemma="de" type="SP"> de</w>
        <w lemma="bienandanza" type="NCFS000"> bienandanza</w>
        <w lemma="que" type="PR0CN00"> que</w>
        <w lemma="ni" type="RN"> ni</w>
        <w lemma="al" type="SP"> al</w>
        <w lemma="yelo" type="NCMS000"> yelo</w>
    </l>
```

https://github.com/bncolorado/CorpusGeneralPoesiaLiricaCastellanaDelSigloDeOro/blob/master/FrayLuisDeLeon/frayLuisDeLeon_AlmaRegionLuciente.xml

Anotación multinivel en CLS - Caso 1

```
 <l met="1|-|-|-|-|6|-|-|-|10|-|" n="2">
        <!--prado de bienandanza que ni al yelo-->
        <w lemma="prado" type="NCMS000">pra|do|</w>
        <w lemma="de" type="SP">de|</w>
        <w lemma="bienandanza"
type="NCFS000">bien|an|dan|za|</w>
        <w lemma="que" type="PR0CN00">que|</w>
        <w lemma="ni" type="RN">ni</w>
        <w lemma="al" type="SP">al|</w>
        <w lemma="yelo" type="NCMS000">ye|lo|</w>
     </l>
```

https://github.com/bncolorado/CorpusGeneralPoesiaLiricaCastellanaDelSigloDeOro/blob/master/FrayLuisDeLeon/frayLuisDeLeon_AlmaRegionLuciente.xml

Anotación multinivel en CLS - Caso de estudio 2

El **hipérbaton** como rasgo estilístico en la poesía del Siglo de Oro.

Niveles implicados:

1. Métrico-rítmico (y rima).
2. Léxico-categorial.
3. Sintáctico.

Caso de estudio 2 - Hipérbaton

```
<l xml:id="LopeVega.1063.1" n="1"> Un soneto me manda hacer
Violante.</l>
```

source.xml

Caso de estudio 2 - Hipérbaton

```
<l source="#LopeVega.1063.1" met="--+--+-+-+-" real="--+--+-+-+-"
enjamb="no">
    <seg n="1" type="syllable" subtype="0"> Un</seg>
    <w n="1"/>
    <seg n="2" type="syllable" subtype="0"> so</seg>
    <seg n="3" type="syllable" subtype="3"> ne</seg>
    <seg n="4" type="syllable" subtype="0"> to</seg>
    <w n="2"/><pause type="1"/>
    <seg n="5" type="syllable" subtype="0"> me</seg>
    <w n="3"/>
    <seg n="6" type="syllable" subtype="3"> man</seg>
    <seg n="7" type="syllable" subtype="0"> da<w n="4"/>ha</seg>
    <seg n="8" type="syllable" subtype="0"> cer</seg>
    <w n="5"/>
    <seg n="9" type="syllable" subtype="0"> Vio</seg>
    <seg n="10" type="syllable" subtype="3"> lan</seg>
    <seg n="11" type="syllable" subtype="0"> te</seg>
    <w n="6"/><pause type="3"/>
</l>
```

metre.xml

Caso de estudio 2 - Hipérbaton

```
<l source="#LopeVega.1063.1">
    <w n=”1” lemma="uno" pos="DI0MS0"> Un</w>
    <w n=”2” lemma="soneto" pos="NCMS000"> soneto</w>
    <w n=”3” lemma="me" pos="PP1CS00"> me</w>
    <w n=”4” lemma="mandar" pos="VMIP3S0"> manda</w>
    <w n=”5” lemma="hacer" pos="VMN0000"> hacer</w>
    <w n=”6” lemma="Violante" pos="NP00000"> Violante</w>
</l>
```

PoS.xml

Caso de estudio 2 - Hipérbaton

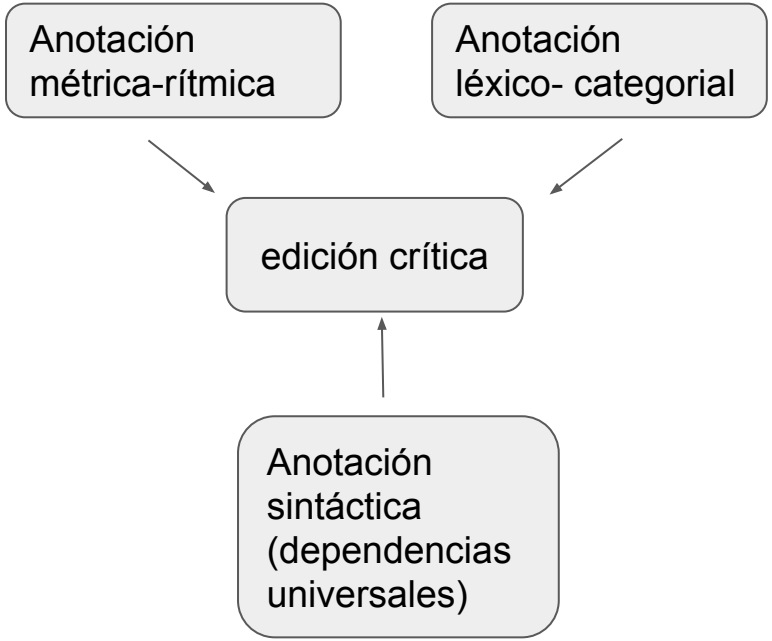
```
1 Un    uno     DI0MS0  DI  - - - - 2 spec    - -
2 soneto   soneto   NCMS000 NC  - - - - 5 obj    - -
3 me    me      PP1CS00 PP  - - - - 4 obl     - -
4 manda mandar   VMIP3S0 VMI - - - - 0 sentence - -
5 hacer hacer   VMN0000 VMN - - - - 4 obj     - -
6 Violante violante NCMS000 NC  - - - - 4 suj    - -
7 .     .       Fp      Fp  - - - - 4 f    - -
```

syntax_conll.txt

Anotación multinivel en CLS - Idea general

Cuatro ficheros contienen toda la información:

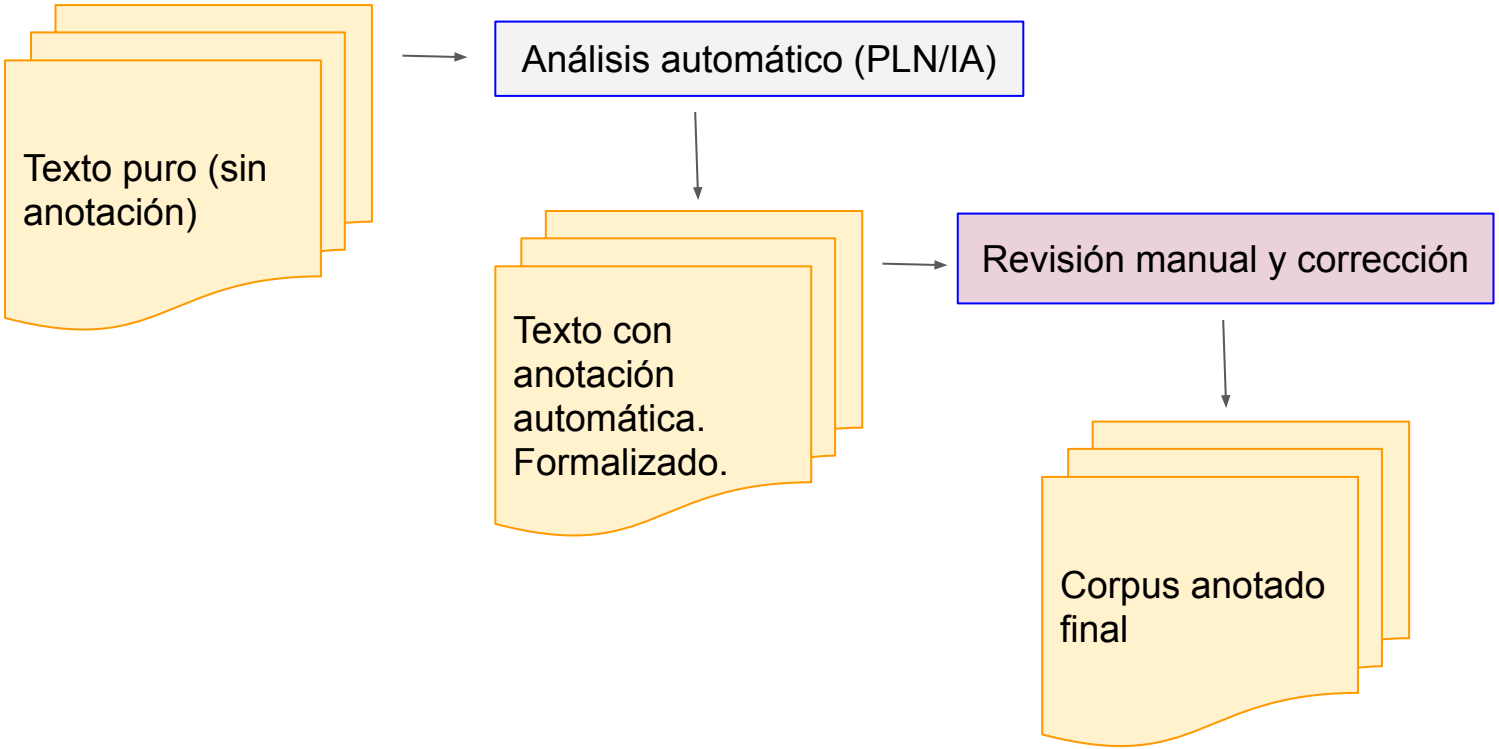
- source.txt: texto
- metre.xml: métrica
- pos.xml: cat. gramaticales
- syntax.conll: con información de posición de palabras y relaciones sintácticas.



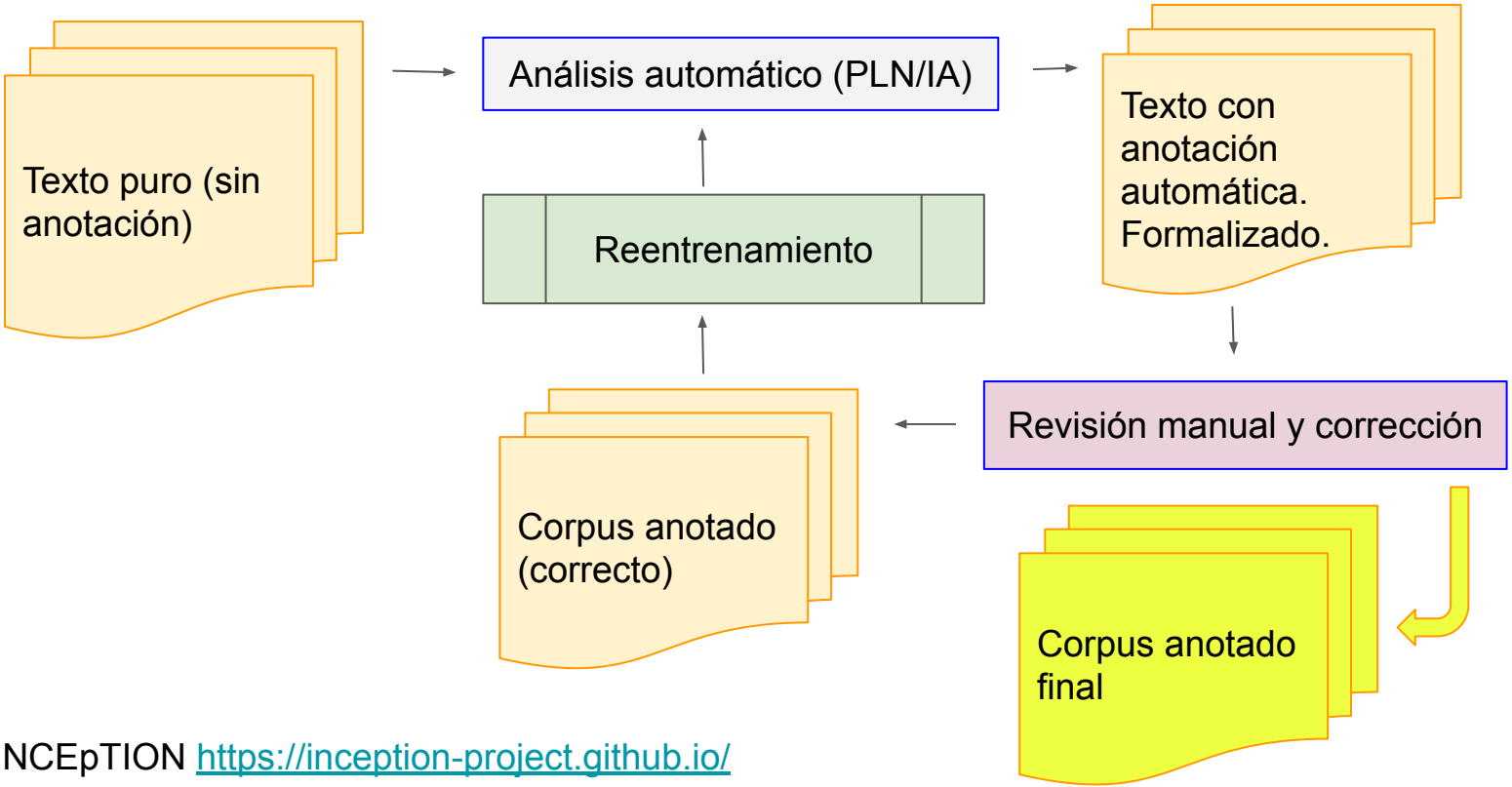
Anotación de un corpus literario

- ¿Qué información anotar?
- Formalización de la información.
- **Métodos de anotación.**
  - Automática (Procesamiento del Lenguaje Natural).
  - Manual.
  - **Semi-automática: *active learning*.**

Anotación semiautomática



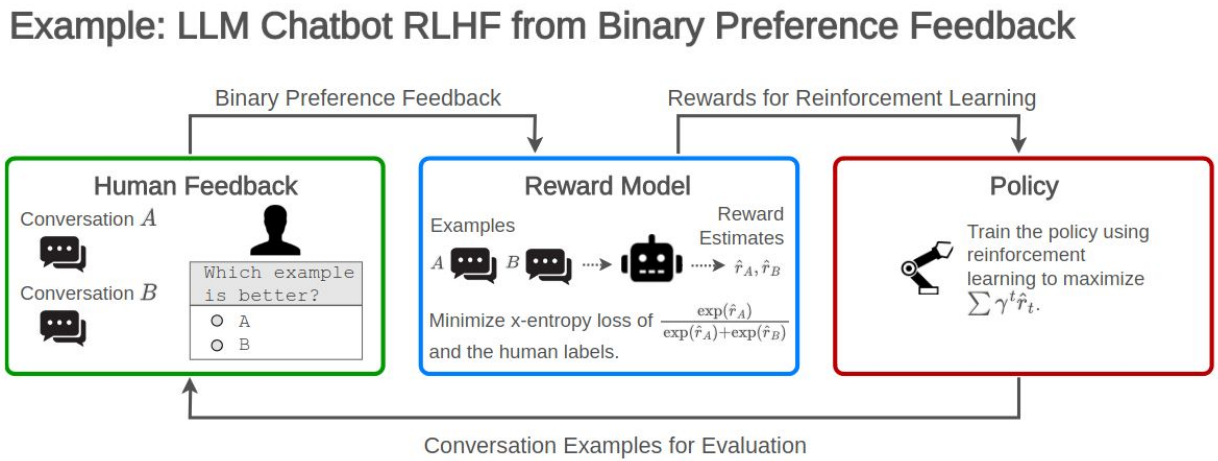
Active learning



Vid INCEpTION https://inception-project.github.io/

Reinforcement Learning from Human Feedback

- “ prominent technique to adapt machine learning models to difficult-to-specify goals” (Casper et al 2023)
- Método para introducir anotación manual en modelos neuronales basados en aprendizaje por refuerzo (*reinforcement learning*).



Conclusiones

- Estudios literarios computacionales no es solo aplicar herramientas de análisis automático al texto literario.
- **Edición crítica digital enriquecida**: más allá de la fijación del texto.
  - Anotación: fijación de (un) análisis / interpretación de fenómenos literarios.
  - Campo para el análisis literario “tradicional”.
  - Forma de “enseñar” a la máquina a **analizar y extraer rasgos literarios del texto**.

Bibliografía

1. Biber (1993) “Representativeness in corpus desing” *Literary and Linguistic Computing* 19, 219-241
2. Casper et al (2023) “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback” pre-print https://arxiv.org/pdf/2307.15217.pdf
3. Egbert, Jesse (2019) «Corpus Design and Representativeness». En *Multi-Dimensional Analysis*, Tony Berber Sardinha y Marcia Veirano Pinto (eds). London, New York: Bloomsbury Academics, 2019.
4. García Berrio, A. (1978) “Lingüística del texto y tipología lírica (La tradición textual como contexto)”, *Rev. Española de Lingüíst.* 8, 19–75.
5. Hatzel, Hans Ole, Haimo Stiemer, Chris Biemann, y Evelyn Gius (2023) “Machine Learning in Computational Literary Studies”. *It - Information Technology* https://doi.org/10.1515/itit-2023-0041.
6. Jauralde Pou, Pablo (2020) *Métrica española*. Madrid: Cátedra.
7. Manero Sorolla, María del Pilar (1990) *Imágenes petrarquistas en la lírica española del renacimiento, repertorio*, Barcelona, PPU.
8. Navarro Colorado, Borja; María Ribes Lafoz and Noelia Sánchez (2016) "Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation", *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, 23-28 May 2016, Portorož (Slovenia) [PDF].
9. Navarro Colorado, B. (2019) "Por un análisis distante y profundo: un corpus piloto de la poesía lírica castellana del Siglo de Oro", Revista de poética medieval, (33), 51-76. https://recyt.fecyt.es/index.php/revpm/article/view/69109
10. Navarro Tomás, Tomás (1956, 1995) *Métrica española*. Barcelona: Labor.
11. Odebrecht, Carolin; Burnard, Lou; Navarro-Colorado, Borja; Eder, Maciej; Schöch, Christof (2019) "The European Literary Text Collection (ELTeC)", Digital Humanities conference (DH 2019), Utrecht, 9-12 July.
12. Quilis, Antonio (1993) *Métrica española*. Barcelona: Ariel.
13. Rivers, E. L. (1993). *El Soneto Español en el Siglo de Oro*. Madrid, Akal.
14. Romojaro, Rosa (2019) *Lope de Vega y la teoría de las funciones del mito*, Barcelona, Anthropos. 527 páginas. ISBN: 978-84-16421-45-9
15. De Rose, Steven (2004) "Markup overlap: a review and a horse". *Proceedings of Extreme Markup Language*. http://www.mulberrytech.com/Extreme/Proceedings/html/2004/DeRose01/EML2004DeRose01.html
16. …

APÉNDICE

Representatividad del corpus

Representatividad del corpus

Corpus aleatorios vs. no aleatorios (Egbert 2019):

- Aleatorio: selección totalmente aleatoria de los textos a partir de la **totalidad** de la población.
  - Permite hacer **generalizaciones a partir de la muestra**.
- No aleatorios o de conveniencia:
  - Selección de textos según las necesidades del estudio.
  - Las conclusiones no son generalizables más allá del corpus.
  - **Corpus balanceados**: selección de los textos en función de determinadas categorías procurando que la cantidad de textos por categoría quede compensada.

Caso 1

Corpus del soneto del Siglo de Oro

https://github.com/bncolorado/CorpusSonetosSigloDeOro

- Todo soneto digitalizado entre Garcilaso de la Vega (+/- 1543) y sor Juana Inés de la Cruz († 1695).
- Mínimo 10 sonetos por autor.
- Representatividad:
  - Garcilaso de la Vega: 38 sonetos conocidos.
  - Lope de Vega: 1382 sonetos aprox.
  - …

Representatividad del corpus

> *Representativeness refers to the extent to which a sample includes the full range of variability in a population* (Biber 1993)

Dos tipos de “representatividad”:

- Representatividad del campo objeto de estudio (*target domain*).
- Representatividad del fenómeno lingüístico o literario.

Ejemplo: estudio de las formas metafóricas en la obra en prosa de Quevedo.

Corpus balanceados: criterios de selección

- Definir y justificar qué criterios se aplican para seleccionar los textos del corpus.
- Establecer la cantidad de textos necesaria por cada criterio.
- Criterios:
  - Género literario: lírica, drama, novela…
  - Idioma.
  - Autoría: sexo, año de nacimiento, procedencia...
  - Edición de la obra: primera, última supervisada por el autor...
  - Periodo, fechas de publicación…
  - Determinados rasgos literarios: temas y subtemas, métrica, estilo, metáforas y tropos, motivos…
- ¿El tamaño importa?

Caso 2

Corpus ELTeC (*European Literary Text Collection*)[1]:

- Corpus de novela europea (1840-1920)
- Objetivo:
  - *build a multilingual European Literary Text Collection (...) containing around 2,500 full-text novels in at least **10 different languages**, permitting **to test methods and compare results across national traditions**.*

1. Proyecto Distant Reading for European Literary History (COST Action CA16204) 2017-2022.

Caso 2 (corpus ELTeC)

Criterios de selección:

100 novelas por cada idioma seleccionados según los siguientes criterios:

- Género literario: novela (prosa narrativa ficcional).
- Lugar de publicación: Europa.
- Diversidad: una novela por autor:
  - Pero once autores podrían estar representados con dos o tres novelas.
- No traducciones.
- A ser posible, la primera edición en formato libro.
- …

Caso 2 (corpus ELTeC)

Criterios de selección:

- Fecha de publicación de la primera edición: 1840 - 1920.
  - T1: 1840-1859
  - T2: 1860-1879
  - T3: 1880-1899
  - T4: 1900-1920

  Compensando: 25 novelas por cada periodo.

Caso 2 (corpus ELTeC)

Criterios de selección:

- Tamaño.
  - *short* (10kv~50k word tokens)
  - *medium* (50kv~100k word tokens)
  - *long* (>100k word tokens)

  Compensado: cada tamaño debe tener mínimo 20 novelas.

Caso 2 (corpus ELTeC)

Criterios de selección:

- Sexo / género del autor.
  - Al menos 10% novelas escritas por mujeres, hasta el 50%.

Caso 2 (corpus ELTeC)

Criterios de selección:

- “Canonicidad”: cantidad de reimpresiones
  - Al menos 30 novelas con solo una edición (la original): canonicidad baja.
  - Al menos 30 novelas con dos o más ediciones modernas en papel (1970-2027): canonicidad alta.

Caso 2 (corpus ELTeC)

Criterios de selección (resumen):

- 100 novelas por idioma.
- Publicadas en Europa.
- No traducciones.
- Una novela por autor.
- A ser posible, la primera edición en formato libro.
- Compensado en fecha de publicación (25 novelas por cada periodo de 20 años)
- Compensado en tamaño (*sort, medium, large*)
- Compensado en sexo del auto (10~50%)
- Compensado en cantidad de reimpresiones.

Caso 2 (corpus ELTeC)

Situación actual

https://distantreading.github.io/ELTeC/index.html

✣★✳❄✣✍❋❑❒❄				AUTHORSHIP				LENGTH			TIME SLOT					REPRINT COUNT		
Language	Last update	Texts	Words	Male	Female	1-title	3-title	Short	Medium	Long	1840-59	1860-79	1880-99	1900-20	range	Frequent	Rare	E5C
cze	2021-04-09	100	5621667	88	12	62	6	43	49	8	12	21	39	28	27	1	19	80.00
deu	2022-04-19	100	12738842	67	33	35	9	20	37	43	25	25	25	25	0	48	46	96.92
eng	2022-11-17	100	12227703	49	51	70	10	27	27	46	21	22	31	26	10	32	68	100.00
fra	2022-01-24	100	8712219	66	34	58	10	32	38	30	25	25	25	25	0	44	56	101.54
gsw	2023-03-30	100	6408326	73	27	32	9	45	40	15	6	16	19	59	53	0	0	66.15
hun	2022-01-24	100	6948590	79	21	71	9	47	31	22	22	21	27	30	9	32	67	100.00
pol	2022-06-01	100	8500172	58	42	1	33	33	35	32	8	11	35	46	38	39	61	80.00
por	2022-03-15	100	6799385	83	17	73	9	40	41	19	13	37	19	31	24	26	60	94.62
rom	2022-05-31	100	5951910	79	16	59	9	49	31	20	6	21	25	48	42	24	76	83.08
slv	2022-02-02	100	5682120	89	11	26	5	53	39	8	2	13	36	49	47	48	52	78.46
spa	2022-05-16	100	8737928	78	22	46	10	34	35	31	23	22	29	26	7	46	54	100.00
srp	2022-03-17	100	4931503	92	8	48	11	55	39	6	2	18	40	40	38	38	62	80.77

Recapitulación

- Diferenciar entre representatividad del dominio y la representatividad del hecho literario a analizar.
- A partir de límites precisos (la totalidad del dominio) se puede crear muestras aleatorias. Permite generalización estadística.
- Corpus balanceados: selección de textos a partir de unos criterios determinados, con un cantidad balanceada de textos por cada criterio.