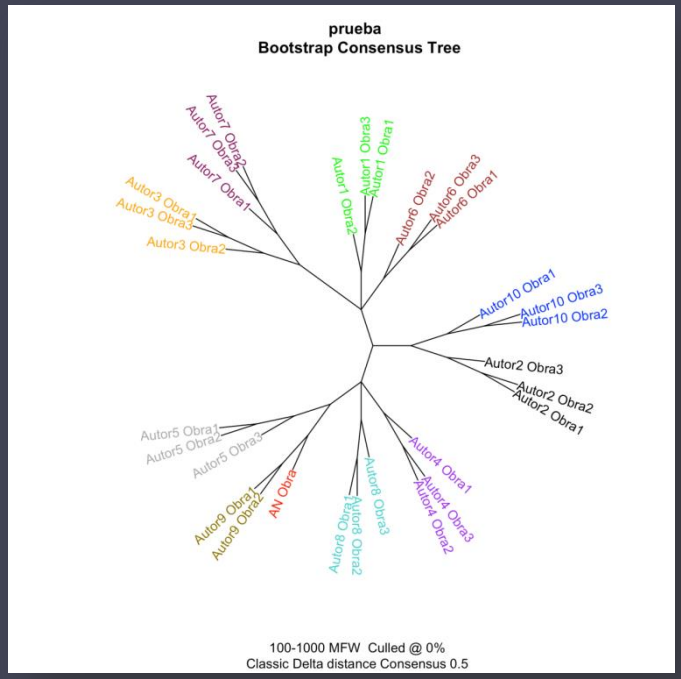
Profundización en stylo. Métricas



Laura Hernández Lorenzo

Humanidades Digitales. Del corpus a la interpretación: Estilometría con R (Universidad de Burgos)

Contenidos

- Métodos supervisados / no supervisados
- Método supervisado: clasificación
- Práctica 2
- Principal Component Analysis
- Práctica 3
- Zeta
- Práctica 4

Clasificación de métodos estilométricos

- Existe una cierta variedad de métodos estilométricos en la actualidad

- "In present-day stylometry, it is common not to limit a study to a single technique, but to compare the output of different methodologies. In the virtual absence of ground truth [...] it is worthwhile to asses the stability of experimental outcomes using different methodologies, which all have their strenghts and weakenesses" (Stover & Kestemont, 2016, p. 654).

Clasificación de métodos estilométricos

- Métodos no supervisados / métodos supervisados

Métodos no supervisados

- Se aplican al corpus las técnicas y se observa de un simple vistazo qué textos son más similares

- Análisis de grupos, árboles de consenso…

Métodos supervisados

- El investigador aplica métodos estilométricos de forma controlada sobre un corpus de textos normalmente de autoría segura para probar la eficacia del análisis
- Métodos de clasificación
- Ventaja: permiten comprobar el grado de fiabilidad que puede esperarse del análisis antes de aplicarlo sobre el texto anónimo.

Práctica 2

- Dentro de la carpeta “novela”, crea una carpeta llamada “classification” y dentro, una con el nombre “primary_set” y otra llamada “secondary_set”
- Mete todos los textos de “corpus” en el “primary_set”. Después extrae algunos de diferentes autores y llévalos al “secondary_set”
- **¡Ojo!** Debe quedar, al menos, un texto por autor en el “primary_set”

Para profundizar

- Cross-validation using the function classify()
- https://computationalstylistics.github.io/blog/cross-validation/

- Performance measures in supervised classification
- https://computationalstylistics.github.io/blog/performance_measures/

Principal Component Analysis (PCA)

- Una de las técnicas más utilizadas para atribución de autoría → destacan los trabajos de Mike Kestemont

- Método procedente de la Estadística

  - Reduce las dimensiones de un conjunto de datos para que sea posible apreciar, en el caso de la Estilometría, qué textos se encuentran más cercanos

PCA

- **¡IMPORTANTE!** Para atribución de autoría, PCA solo debe aplicarse con un conjunto pequeño de autores (idealmente, tres)

- Tiene la ventaja de que obtiene resultados altamente fiables con pequeños corpus

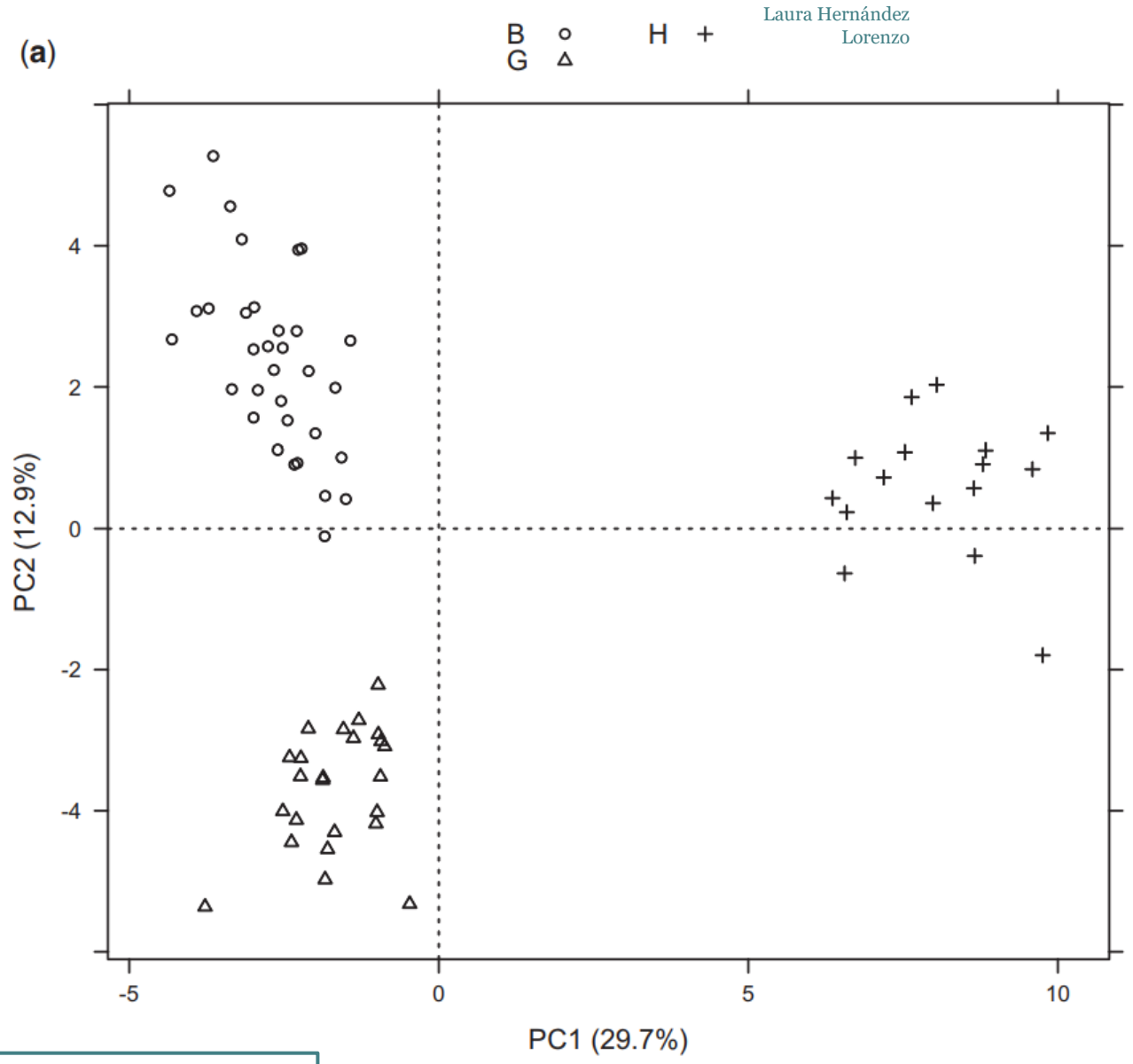
Caso de las *Visio* de Hildegarda de Bingen (Kestemont et al., 2015)

- Caso problemático de autoría
  - *Visio de Sancto Martino* y *Visio ad Guibertum missa*
  - Discusión en la crítica sobre si debían atribuirse a Hildegarda o a su secretario y corrector, Guibert de Gembloux

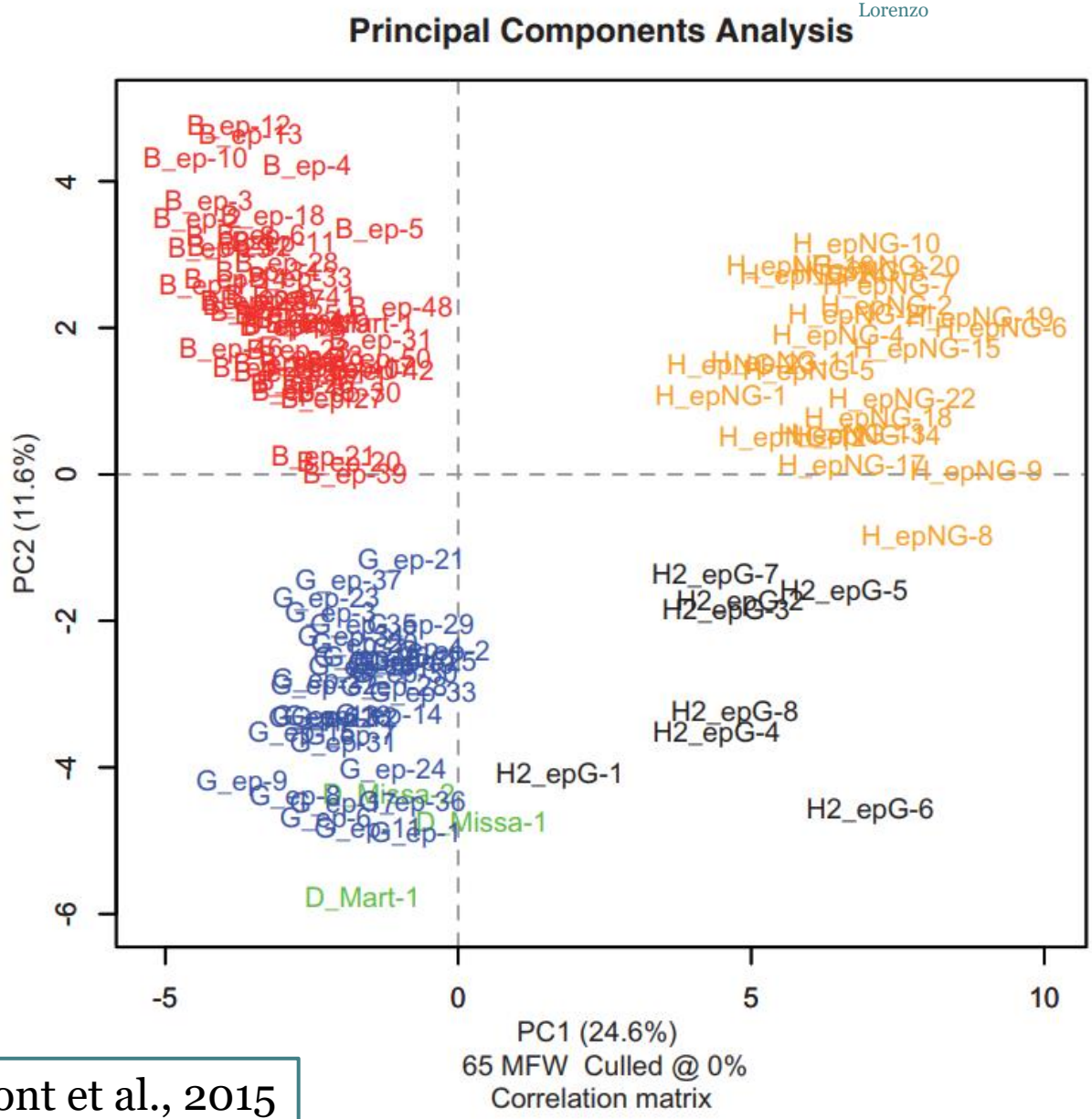
Caso de las *Visio* de Hildegarda de Bingen (Kestemont et al., 2015)

- Kestemont realiza un análisis estilométrico basándose en PCA
- Además de textos de Hildegarda y Guibert, utiliza un autor de control: Bernard de Clairvaux

Documental (en inglés) sobre la investigación en la web personal de Mike Kestemont



Kestemont et al., 2015



Kestemont et al., 2015

Práctica 3

- Dentro de “novela” crea una carpeta que se llame “PCA”, y dentro de esta, una con el nombre “corpus”.
- Copia las obras de Galdós, Pardo Bazán y Clarín a la nueva carpeta “corpus”.
- Cambia el directorio de trabajo en R y ejecuta el comando stylo()

Zeta para comparar textos

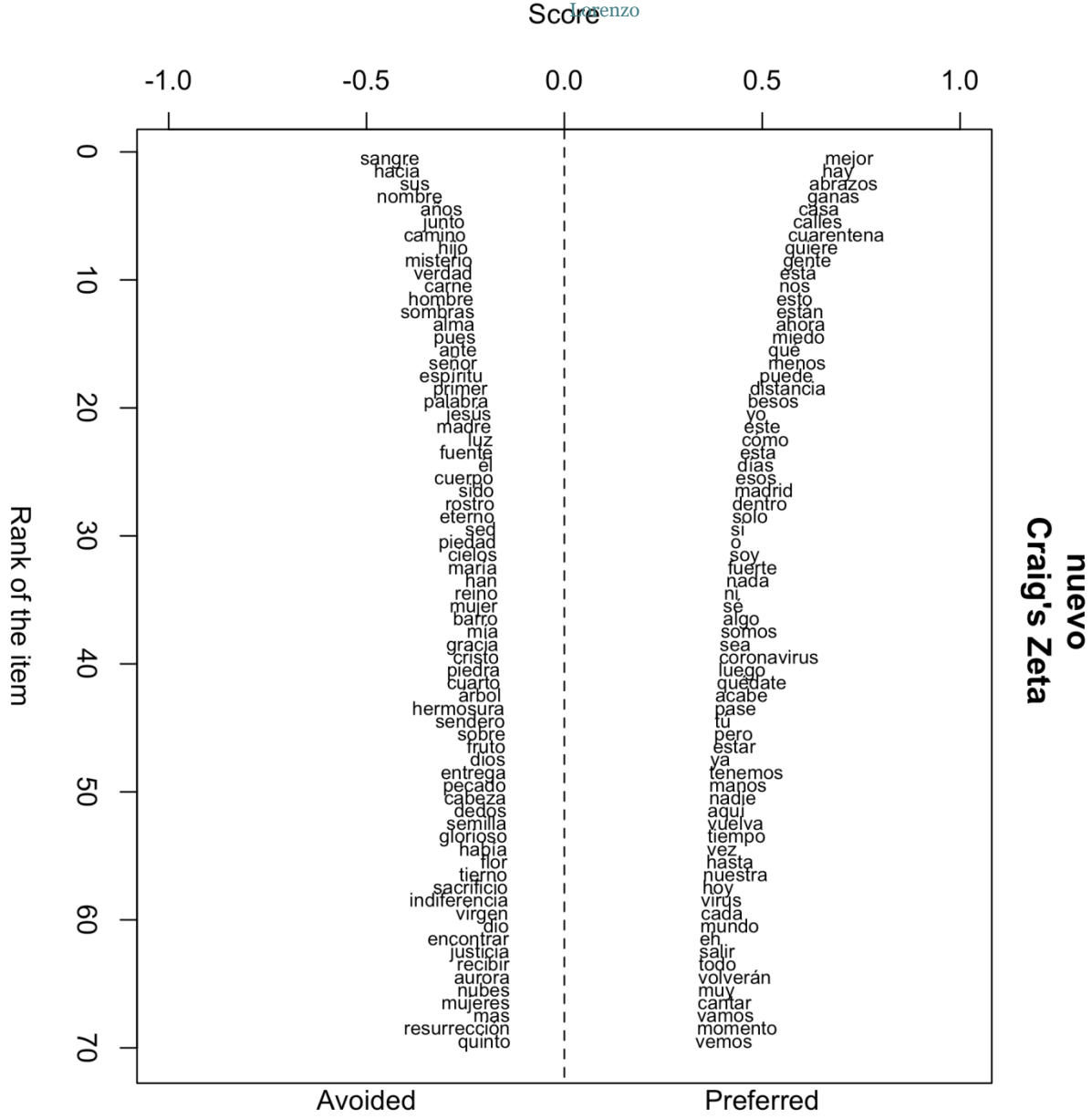
- Zeta → otra de las medidas desarrolladas por Burrows. La variante de Craig es muy utilizada.

- Función oppose() en *stylo* → comparar dos textos y ver las palabras más relevantes de unos frente a otros
  - Relacionado con las palabras clave (*keywords*) que se utilizan en estudios de corpus

Laura Hernández
Lorenzo

Ejemplo de uso de Zeta

¿qué textos crees que he usado?



Práctica 4

- Para aplicar oppose, necesitas otra vez una carpeta llamada “primary_set” y otra llamada “secondary_set”. Crea dentro de la carpeta “novela”, una llamada “zeta”, y dentro dos carpetas con los nombres anteriores.

- Vamos a contrastar las obras de Emilia Pardo Bazán con las de Valera → copia las obras de ella a “primary_set” y las de él a “secondary_set”