What do we need to exploit ELTeC corpus?

Borja Navarro Colorado

University of Alicante

June 2021

We have seen

- how to extract the information annotated in XML
- how to analyze the corpus with basic NLP techniques and extract linguistic or literary data.

Overcoming language barriers

Analyzing European Literature as a whole:

- to find common features and relevant differences of 19th. century European Novels
- to create a map of influences between novels from different languages and countries
- to detect stylistic relations between authors from different countries
- etc.

We need multilingual NLP systems: nowadays a research topic.

A simple proposal

Stylometric relations between Spanish, English and French novels
(from ELTeC corpus).
Process:

1. An inter-lingual representation of each word (nouns, verbs,
   adjectives and adverbs) based on Open Multilingual **WordNet
   synsets**.

   > Tool: UKB WSD algorithm (Agirre and Soroa 2009) that is
   > available in Freeling (Padró et al 2010).

2. To detect stylometric relations with **stylo()**, the R package for
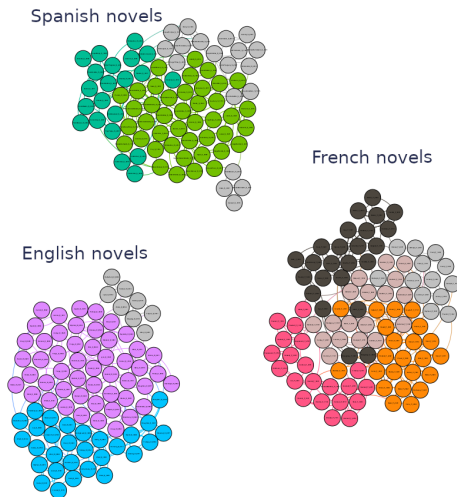   stylometric analysis (Eder et al, 2016).

Experiment 0

Figura: MFW 100; Culling 0; Distance: cosine

Experiment 1

Figura: MFW 1000, start at 50; Culling 100; Distance: cosine

Zoom in/Ampliar imagen

Experiment 1

Some relations:

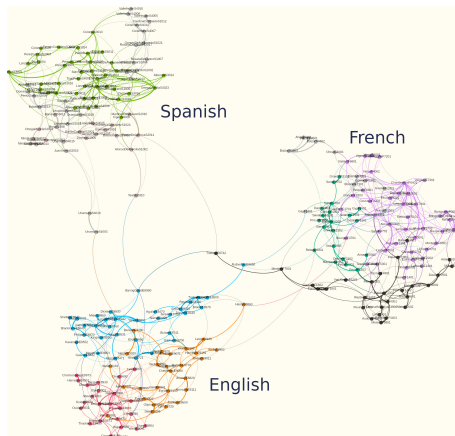| Spanish	English
Blasco Ibañez *La barraca*	Baring-Gould *Domitia*
Pérez Galdos *Fortunata y Jacinta*	Lyall (pseud.) *The Autobiography of a Slander*
Unamuno 1917 *Abel Sánchez*	Arlen (pseud.) 1920 *The London venture*
Llofriu y Sagrera 1872 *El naufragio del grumete*	**William Francis Barry 1887** *The New Antigone* Victoria Cross 1895 *The Woman Who Didn't*

Experiment 2

Figura: Only verbs. MFW 1000, start at 100, culling: 90, cosine distance

Zoom in/Ampliar imagen

Experiment 2

Some relations:

Spanish	English	French
Unamuno *Niebla* 1914	Benson 1919 *Living Alone*	
Avecilla 1852 *La conquista...*	Yeats 1891 *John Sherman...*	
Pastor 1863	**Trollope 1874**	**Mendès 1892**
La Corona	*Harry Heathcote*	*Lucignole*
	Hardy 1886	Bourget 1919
	The Mayor of Casterbridge	*Laurence Albani*

but, what do we need to exploit ELTeC corpus?

Research infrastructures for language and humanities:

- to store the corpus as research resource
- with several tools able to exploit the corpus for diverse purposes.

as CLARIN or DARIAH

Bibliography

- Agirre, E. y A. Soroa (2009) Personalizing PageRank for Word Sense Disambiguation. 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens 2009. `https://www.aclweb.org/anthology/E09-1005/`

- Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. R Journal 8(1): 107-121. `https://journal.r-project.org/archive/2016/RJ-2016-007/index.html`

- Padró, Ll.; S. Reese, E. Agirre y A. Soroa (2010) Semantic Services in FreeLing 2.1: WordNet and UKB. In *Principles, Construction, and Application of Multilingual Wordnets*, eds Bhattacharyya, P., C. Fellbaum y P. Vossen, 99–105, Mumbai: Narosa Publishing House.