ACL 2024
Bangkok, Thailand

Historical Ink: Semantic Shift Detection for 19th Century Spanish

Tony Montes, Laura Manrique-Gómez, Rubén Manrique

Motivation

- Previous attemps to Semantic Shift Detection on LSCDiscovery shared task on Spanish.
- Understanding semantic shifts in 19th-century Spanish and particularly on Latin-American Spanish helps us uncover historical and cultural transformations.
- Our project aims to detect these shifts, providing insights into language evolution during this period with a focus on regional particularities.

LSC DISCOVERY

Data

Project Gutenberg

eBooks

1800-1914

The British Library Books

Digitized Books

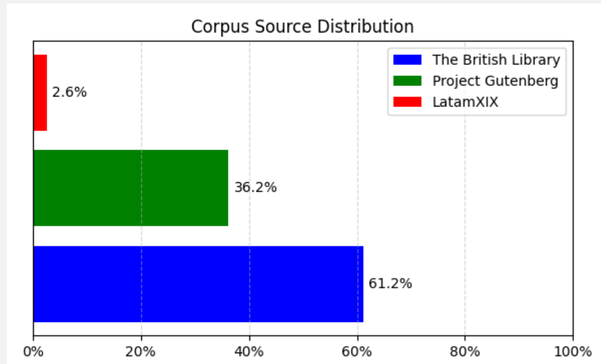
1800-1899

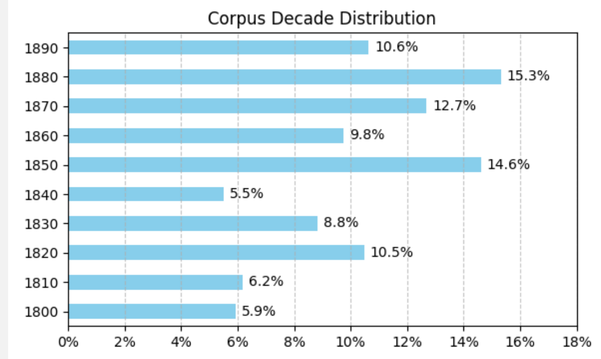
LatamXIX

Latin American Spanish texts from 19th century Newspapers

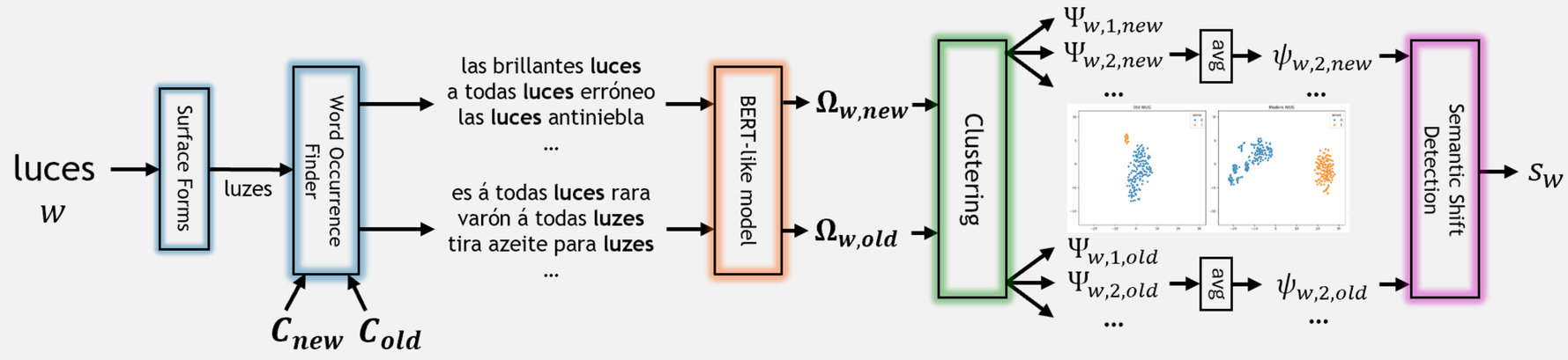
1845-1899

~ 160M tokens





Methodology



LM Training

Trained three BERT-like models on Latin-American portion of the corpus (*LFT*) and some on the whole corpus (*FT*).

LMs used for generating word embeddings

BETO **mBERT** **AlBERT**

Clustering

Worked with two clustering algorithms.
Each cluster/color represents a word meaning

Kmeans (silhouette/inertia optimized)

Affinity Propagation

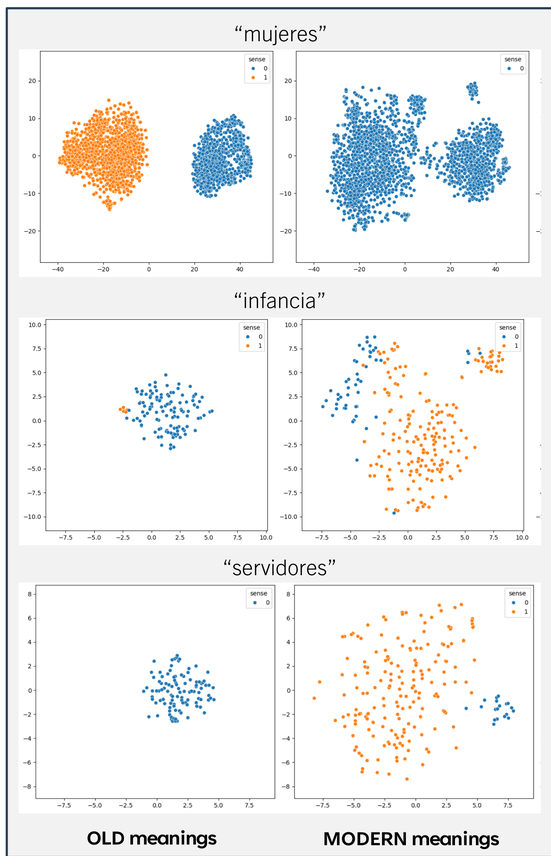
Semantic Shift Detection

Measures how much the meanings of a word changed, and which meanings were gained/lost

Cosine Distance (CD)

Inverted Similarity over Word Protorype (PRT)

Results



lost sense

Evolved from reflecting traditional gender roles to modern gender discourse

gained sense

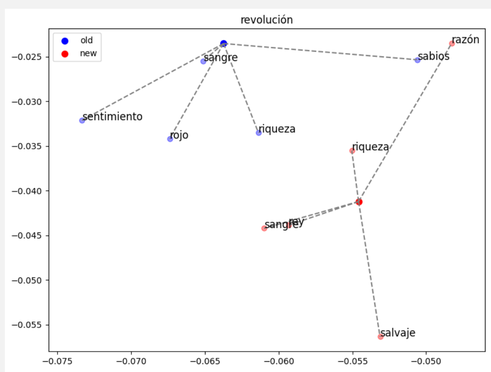
Changed from 'infancy of the nation' to primarily meaning 'childhood'

gained sense

Changed from 'servants' meaning to current tech-related 'servers'

The semantic shift detection was performed on 250^{+} words in Spanish; some of the most insightful observations were the presented words.

The embeddings generated also allow to perform comparisons between different words to compare relations between words.



BETO-cased trained on Latin-American corpus was the best performing model to detect shift of word usages.

The pipeline effectively allows to automate the detection of lost and gained meanings for different words.

	Word	Sense	CD	PRT	gained/lost Sense
AP	Rey	0	0.005	1.005	
	Usurero	0	1.0	∞	**lost**
KMeans	Luces	0	0.012	1.012	
	Luces	1	0.012	1.013	
	Infancia	0	0.017	1.017	
	Infancia	**1**	1.0	∞	**gained**
	Sentimiento	**0**	1.0	∞	**gained**
	Sentimiento	1	0.003	1.003	
	Sublime	**0**	1.0	∞	**lost**
	Sublime	**1**	1.0	∞	**lost**
	Sublime	**2**	1.0	∞	**lost**
	Servidores	0	0.043	1.045	
	Servidores	**1**	1.0	∞	**gained**

Universidad de los Andes Colombia | Departamento de Ingeniería de Sistemas y Computación

Departamento de **Historia y Geografía**
Facultad de Ciencias Sociales

Paper, Code, Data