Exploring Semantic Differences in References to France made by Two Francophone Algerian Authors

Lucas Donat | Faculty Sponsor: Typhaine Leservot
QAC Apprenticeship Summer 2025

Introduction

- In postcolonial literature, that is literature written by authors from colonized and/or formerly colonized territories, artists have long created art that revisits past colonial violence inflicted on their home nations by the colonizers. For authors from North Africa living in North Africa and France from the 1950s until the end of the 20th century, this has meant focusing on revealing living conditions under France's subjugation.
- In recent years, developments in natural language processing have led to the creation of text embedding models, which capture much more nuance than previous computational text analysis techniques.[1]

Research Questions

Considering the aforementioned evolving relationship to France in the 21st century, this project aims to explore the ways in which text embedding analysis can provide fruitful information for literary analysis[2], leading us to ask:

- How do sentiments about France differ by text and by author in this sample?
- Are there any productive ways we can group these texts to reveal semantic similarities?

Methods

Sample

- This study utilizes text data from 8 fiction novels written from 10/20/2006 - 3/7/2012 by 2 Algerian authors who live in Canada, Karim Akouche and Salah Benlabed.
  - Quote boundaries were drawn 1 sentence before and 1 sentence after keyword ("France") detection.
- Relevant quotes (n = 101) were then processed to ensure homogeneity.
  - Quotes were attributed to each author (Akouche = 41, Benlabed = 60). Contextually unimportant punctuation was also removed at this step.

Measures

- Quote text embeddings were measured using the 2024 spaCy Large French Universal Dependency Library (fr_core_news_lg 2024).
  - The Large French Universal Dependency Library is a freely available Python library of 500,000 words, each represented by embedding vectors of length 300.
- Cluster membership was determined via a constrained k-means process on the principal components of the text embeddings, which were mapped from 23 to 2 dimensions through t-distributed Stochastic Neighbor Embedding (t-SNE).

Preliminary Results

Fig. 1

- Principal Component Analysis (PCA) showed that 23 components were needed to explain at least 90% of the variance in the text embeddings for all quotes.
  - Embeddings were truncated to length 100 to perform PCA so as to avoid overfitting the data.

Fig. 2

- An iterative constrained k-means function determined 2 clusters (of sizes 50 and 51, respectively) to be the most viable quote arrangement across 10,000 trials (n = 1766, 17.66%).
  - Chi-square difference test showed a significant relationship between author and cluster membership (p = 0.00886).
  - Additionally, chi-square showed a significant relationship between the novel from which the quote was extracted and cluster membership (p = 0.01255).

Cluster	Karim Akouche	Salah Benlabed
1	29	21
2	12	39

***Fig. 1*, Principal components vs. explained variance**

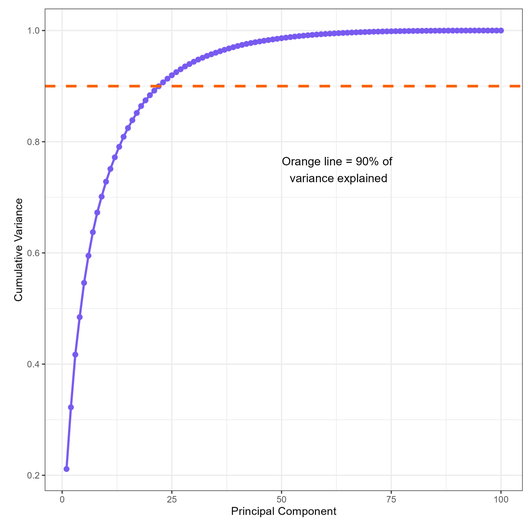
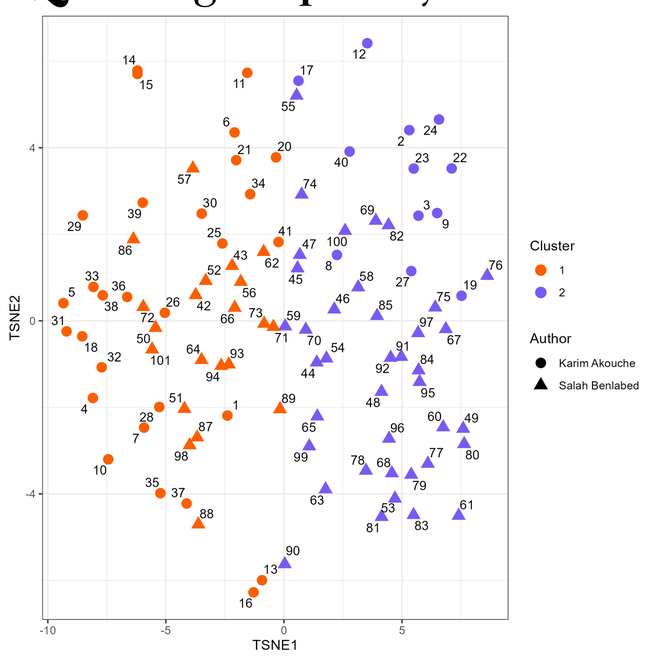


***Fig. 2*, Quotes grouped by t-SNE clusters**



Discussion

- Significant cluster membership differences by author could suggest generational differences in attitudes towards France; that said, the presence of quotes from both authors in each cluster could show some semantic correlation in references to France.
- Analyzing these clusters provided interesting insights: passages in cluster 1 tended to negatively mention France in the past tense, both in political and personal contexts.
  - Texts in cluster 2 tended to more neutrally mention France in contexts of possible future scenarios.
  - These tendencies were consistent regardless of author or parent text.
- Further research could include hierarchical linear modeling to analyze embedding variance between/among authors.
  - Additionally, greedy k-means clustering iteration could be used to find text grouping patterns across 10,000 trials.

References

1. Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.*
2. Underwood, T. (2016). "Distant Reading and Recent Intellectual History." In M. K. Gold & L. F. Klein (Eds.), *Debates in the Digital Humanities 2016* (pp. 530–533). University of Minnesota Press. https://doi.org/10.5749/j.ctt1cn6thb.47

I would like to thank Professor Typhaine Leservot, Professor Pavel Oleinikov, and the entire QAC faculty and student cohort for their support throughout this project.