Culturomics of Language Globalization

Samy Bakikerali, Gloria Chen, Bao Dinh, Dakota Ellis, Uyen Le

University of North Carolina at Charlotte

DTSC 4301: Data Science for Social Good

Dr. Mirsad Hadžikadić and Dr. Ted Carmichael

January 31, 2025

Description of your problem

In this research, we are taking a data driven approach to analyze and uncover patterns in how languages have changed and evolved over time. Many factors such as trade, colonization, or the internet have reshaped the evolution of languages, or even developed hybrid languages.

Relevant theories

- **Language Globalization Theory:** The process in which a language like english becomes dominant and spoken across the world due to changes in trade, culture, and technology
- **Corpus Linguistics:** the method of analyzing large datasets of naturally occurring language to find patterns and associations.
- **Diffusionism:** Anthropological school of thought that describes how cultural traits radiate out from a central origin but influences cultures around the world.

Key Terms

- **Pidgin:** A simplified, makeshift language that develops for communication between speakers of different native languages, with limited grammar and vocabulary. No native speakers.
- **Creole:** A fully developed natural language that evolves from a pidgin when it becomes the first language of a community, gaining complex grammar and vocabulary.
- **Carcinisation of Language:** Different languages developing similar grammar or vocabulary due to repeated exposure to dominant linguistic trends (e.g., English loanwords spreading globally).
- **Language and Cultural Homogenization**
- **Culturomics:** the application of high-throughput data collection and analysis to the study of human culture.

Datasets

- **Google Books Ngram Viewer Dataset** – Provides word frequency data from millions of digitized books over centuries.
- **Oxford English Dictionary (OED) Historical Thesaurus** – Documents the evolution of English words and meanings.al
- **The Linguistic Data Consortium (LDC)** – Offers historical language corpora, including speech and text data.
- Create our own data by transcribing spoken communication.

Intended solution

Our intended solution is to gain deeper insights into the evolution of language by building upon past research using data driven methodologies. While previous studies have provided valuable historical and qualitative analyses, many have been limited by small scale datasets, lack of computational methods, and insufficient interdisciplinary integration. By leveraging quantitative, evidence based approaches, including big data analytics, natural language processing (NLP), and statistical modeling, we aim to address these gaps.