TALOS ERA CHAIR IN ARTIFICIAL INTELLIGENCE FOR HUMANITIES AND SOCIAL SCIENCES

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Computational Linguistics and Corpus Analysis

Dimitris Bilianos
Postdoctoral Researcher

Funded by the European Union

Horizon ERA Chair TALOS AI4SSH Project funded by the European Commission Grant Agreement n° 101087269, https://cordis.europa.eu/project/id/101087269

Introduction

What this session will cover:

- What is **Computational Linguistics**?
- What is a **Corpus** and why is it useful?
- How can we analyze texts using **AI** and **NLP**?
- **Real-world applications** in Digital Humanities & Social Sciences

What is Computational Linguistics?

- Definition: Computational Linguistics (CL) is **the study of language using computers** (Hausser, 2014).
- It combines **linguistics**, **AI**, and **computer science** to help machines process and understand human language.

Key Questions in CL:

- How can a computer read and understand text?
- How do we teach computers to recognize meaning, grammar, and structure in language?

CL in everyday life

- Examples of CL in everyday life:

Machine Translation (e.g., Google Translate) (Koehn, 2009)

Speech Recognition (e.g., Siri, Alexa) (Yu & Deng, 2016)

Text Analysis & Chatbots (e.g., AI customer support) (Jockers & Thalken, 2020)

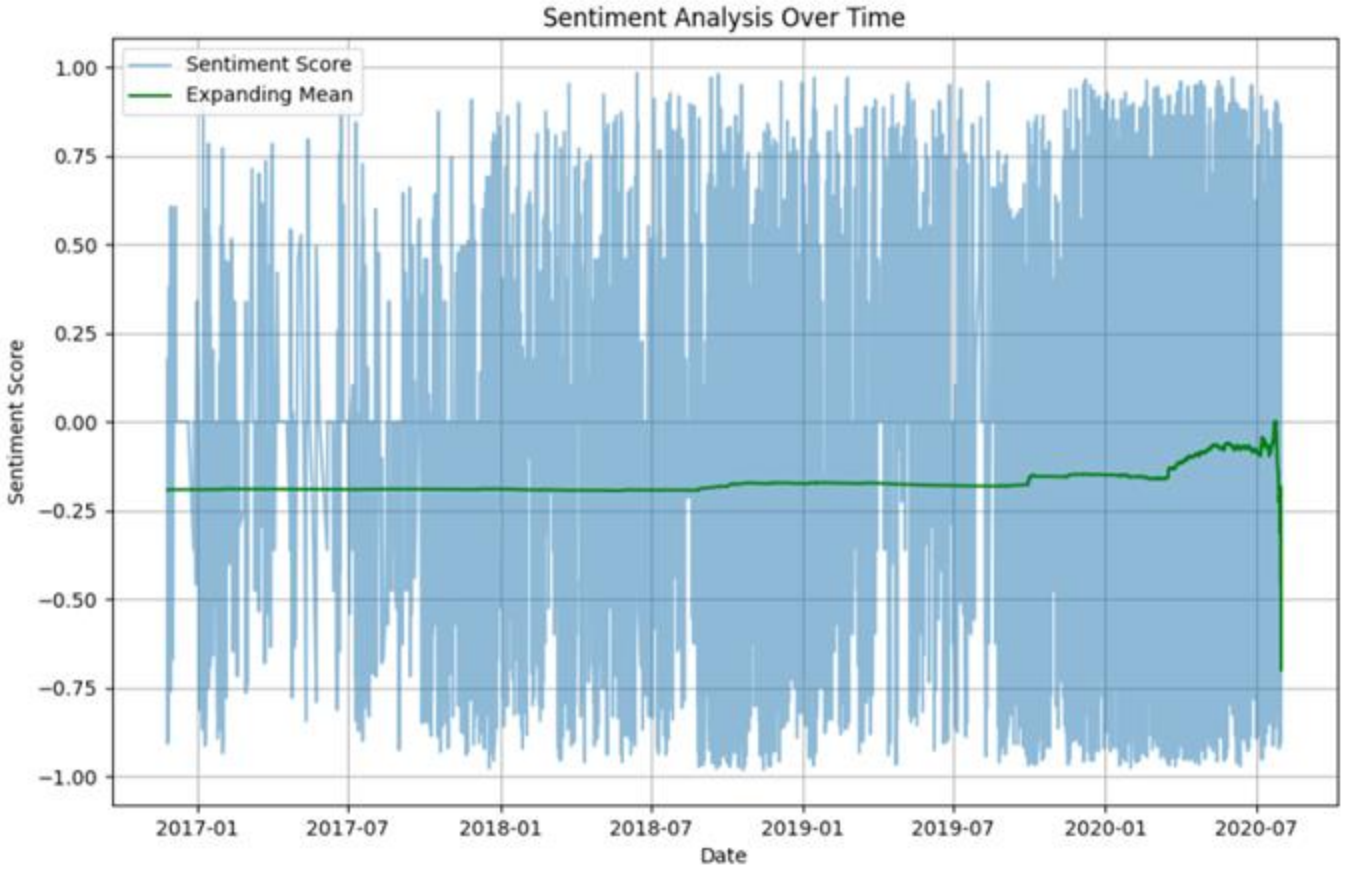
- Why does it matter?

CL helps us analyze large amounts of text automatically, making it essential for Digital Humanities, Social Sciences, and AI applications.

How is CL used in research? - Some examples

- **Topic Modeling** - Discover hidden themes in large text collections (Blei, Ng & Jordan, 2003).
- **Sentiment Analysis** - Detect emotions in texts; analyze public opinion in social media or political speeches (Liu, 2015).
- **Named Entity Recognition** - Identify names of people, places, and organizations in a text (Mohit, 2014).
- **Authorship Attribution** - Determine who wrote a text based on linguistic patterns, e.g., analyzing anonymous or disputed works (Stamatatos, 2009).

Sentiment Analysis



Named Entity Recognition

(...) “ἀναβὰς δὲ εἰς ᾽Ιλιον ἔθυσε τῇ Ἀθηνᾷ καὶ τοῖς ἥρωσιν ἔσπεισε”. → [“᾽Ιλιον”/LOC, “Ἀθηνᾷ”/PER]

Large Language Models

- LLMs are advanced AI models trained on massive amounts of text to understand language and generate human-like language.
- They use deep learning to process text at an unprecedented scale (Achiam et al., 2023).

How do LLMs work?

- Trained on billions of words from books, websites, and articles.
- Predict the most likely next word in a sentence (but in a highly sophisticated way).
- Learn grammar, context, and even subtle nuances of language. (Devlin et al., 2019)

LLMs in Action

- Chatbots & Virtual Assistants
- Text Summarization
- Automatic Translation
- Creative & Academic Writing (eg. Assisting in drafting articles, essays, and stories)

→ LLMs are **transforming how we analyze and generate text**, making them essential for Digital Humanities, historical text analysis, and corpus-based research. However, they also raise **ethical concerns**, such as bias, misinformation, and explainability, which must be critically examined. (Jiao et al., 2024)

The Role of Corpora in Computational Linguistics

A corpus is a large, structured collection of texts that serves as the foundation for training AI models, studying language, and developing NLP applications. (McEnery, 2019)

- **LLMs are trained on huge corpora** containing books, websites, and academic texts.
- Corpora help AI models learn grammar, syntax, and meaning from **real-world language data**.
- Quality of training data = quality of AI output → Biases in corpora can affect AI-generated text.

Corpora for Computational Linguistics, Social Sciences and the Humanities

- Historical & Literary Corpora can be used to study **language evolution** and **authorship attribution**.
- Social Media & News Corpora → Help **analyze public opinion** and sentiment trends.
- Multilingual Corpora → Essential for **machine translation** and **cross-linguistic analysis**.

→ Corpora are **essential for both training AI and conducting research** in linguistics, humanities, and social sciences.

Summary & Conclusion

- CL combines language study and AI to help computers **process, analyze,** and **generate** human language.
- LLMs are **trained on vast amounts of text**; they power applications like chatbots, translation tools, and text analysis.
- Corpora are **large collections of texts** that allow us to study language, meaning, and patterns at scale; Corpus Analysis helps researchers find insights in language and beyond.

→ Computational Linguistics and Corpus Analysis are powerful tools for linguists, historians, social scientists, and AI researchers. They help us understand language on a deeper level, analyze vast amounts of text efficiently, and build AI models that interact with human language.