**PROJECT AIMS**

**Samtla** (Search And Mining Tools with Linguistic Analysis) is an online integrated research environment designed in collaboration with historians and linguists to facilitate the study of digitised texts written in any language.

It currently supports the research of a collection of Aramaic Incantation texts from late antiquity written in Jewish Aramaic, Judeo-Arabic, Syriac, and Mandaic dialects (figure 1). Another version supports the King James Bible in English as a proof of concept of its language independent design.

In contrast to standard search engines and text mining systems that rely on the bag-of-words representation of text, **Samtla** provides the retrieval and discovery of fuzzy text patterns/motifs (aka “formulae” to historians), which is achieved through applying a character-based n-gram statistical language model built on top of a powerful generalised suffix tree data structure.

**CURRENT STATUS**

**Browsing**

Multiple entry points to the archive through faceted navigation using, ontologies, meta-data (author, year, notes, publications), and document features (document length, named entities).

- Documents are presented together with relevant meta-data in any format or language, including images and video where available.

**Search**

- Built on Statistical Language Models and character-based suffix trees. Provides users with tolerant sequence matching and a probabilistic approach to scoring search results (figure 2).

**Query recommendation**

- Users are presented with alternative queries i.e. phonological differences represented by orthographic variations (different spellings).

**Recommended documents**

- The system recommends documents to users, which are semantically similar to the one they are viewing.

**Document comparison**

- Document comparison tool provides a finer-grained comparison by allowing users to analyse common sequences shared by pairs of documents (figure 3).

**Community support**

- Enables users to explore parts of the archive, which they may not have considered as part of their research agenda, by leveraging user activity data.

**CURRENT WORK**

**Named entity recognition**

- Provides alternative visualisation layers over the documents with Named Entity (people, places, and things). Encyclopedic knowledge and Google Maps are used to provide a deeper understanding of the entities and their role in the document when hovering over the links in the document.

**FUTURE WORK**

- We are working with new user groups to develop tools tailored to their research needs.

---

**Figure 1** – Samtla for Armaic Magic Bowls from Late Antiquity (document view).

**Figure 2** – Samtla search is flexible providing full and partial query matching (King James Bible).

**Figure 3** – Document comparison tool enables users to compare documents based on both large and small shared sequences.

---

**Samtla**

A language independent research environment designed to unlock digital document collections through search and comparison tools.

**Project Investigators**

Prof. Mark Levene
Dr Dell Zhang
Martyn Harris

Birkbeck University College London

Dr Dan Levene
Southampton University

**Webste and Publications**

www.samtla.com

**Keywords**

Digital Humanities
Information Retrieval
Statistical Language Modeling
Text Mining

---

**Figure 2**

Kings chapter 12

Malachi chapter 27

Matthew chapter 15

- He received it.
- and they that be disobedient, and nourished them:
- if they be not the sons of God, come from them.

- 41. Likewise also the chief priests mocking him, with the scribes, and elders, said, 42. He hath said, I am Christ the son of God, 43. and the chief priests, and all the council did likewise.

- 10. And in the same manner the King James Bible.

- 30. save the blind, and come down from the cross.

- 33. And in the ninth hour there was darkness over the whole land until the ninth hour.

- 44. And as the ninth hour was come, there was darkness over the whole land until the ninth hour.

- 59. And the word of the Lord came to Elijah in the night, saying.

- 68. And the word of the Lord came to Elijah in the night, saying.

- 87. And the word of the Lord came to Elijah in the night, saying.

- 106. And the word of the Lord came to Elijah in the night, saying.