The Journal of Open Source Software

ATAS - Academic Text Analysis System

Alides Baptista Chimin Junior [1]

1 Universidade Estadual do Centro-Oeste (UNICENTRO)

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Richard Littauer

Reviewers:

- @DiegoAscanio
- @rafaelanchieta
- @felipemaiapolo

Submitted: 25 February 2025
Published: unpublished

Abstract

The ATAS - Academic Text Analysis System (Brazilian Portuguese SATA - Sistema de Análise de Textos Acadêmicos) is an open-source software developed to assist researchers in textual content analysis. Inspired by the methodology of Bardin (2011), ATAS facilitates the extraction, filtering, and statistical analysis of academic texts, enabling the identification of semantic and linguistic patterns. The software is particularly useful for researchers in the Humanities and Social Sciences, providing tools for analyzing keywords, bigrams, lexical categories, and other quantitative metrics of textual analysis. It integrates natural language processing (NLP) and allows data export to software such as Bastian et al. (2009) for semantic network analysis. We emphasize that the tools are in Portuguese, as they were developed by a Brazilian researcher and are being used by the GEPES and GETE research groups.

Statement of Need

In the Brazilian Humanities and Social Sciences research context, many scholars still rely on manual or semi-manual methods to process and analyze large textual corpora. According to Metzler et al. (2016), barriers such as limited programming skills, lack of access to adequate infrastructure, and scarce training opportunities hinder the adoption of computational approaches in these fields. As a result, essential tasks like identifying thematic patterns, generating bigram networks, classifying authors by gender, and calculating lexical statistics often become labor-intensive and error-prone. The ATAS – Academic Text Analysis System addresses these issues by providing an open-source, graphical-interface-based solution that automates these processes without requiring advanced technical expertise. By integrating Natural Language Processing (NLP) capabilities in Brazilian Portuguese, ATAS bridges the gap between sophisticated analytical methods and their practical usability for non-technical researchers. This directly supports more equitable access to computational tools in contexts where language and resource limitations frequently exclude researchers from digital scholarship. Beyond facilitating traditional content analysis, ATAS expands methodological possibilities for examining the relationship between discourse and spatiality. While Geographic Information Systems (GIS) are powerful tools for spatial analysis and thematic cartography, their architecture — based on discrete vector and raster data structures — tends to produce static "snapshots" of space. This structural limitation often prevents them from representing the inherently dynamic, processual, and socially constructed nature of geographic phenomena. As highlighted by Harvey (2005) and Harvey (1980), and further developed by Santos (1996) and Santos (1978), geography extends far beyond cartographic representation. Space is not merely a neutral container of events, but a social, political, and economic construct permeated by power relations, symbolic appropriation, and subjective experiences — dimensions that resist reduction to numerical variables in a GIS database. ATAS offers a methodological alternative by enabling the extraction and analysis of "discursive spatialities" — the ways in which space is constructed, contested, and redefined through language — using spatial statistics and semantic networks directly from textual data. In doing so, it complements rather than replaces cartography, offering researchers in the Humanities and Social Sciences a way to capture spatial meaning

that is processual, contextual, and deeply embedded in discourse.

Features and Usage

1. Text Filtering (Brazilian Portuguese: Filtragem de Texto)

Extracts verbs, adjectives, and nouns from texts, facilitating qualitative analyses.

- **Library used:** `spaCy`
- **How to use:**
  1. Open ATAS and go to the Filter Text option.
  2. Select the `.txt` file to be analyzed.
  3. The system processes the text and saves a new filtered file.

2. Table Conversion (Brazilian Portuguese: Conversão para Tabela)

Generates bigrams from the text and exports the data in CSV format, useful for analysis in Gephi.

- **Library used:** `pandas`
- **How to use:**
  1. Access the Convert Text to Table option.
  2. Choose a `.txt` file.
  3. ATAS generates a CSV file containing the bigrams, ready for network analysis.

3. Gender Identification (Brazilian Portuguese: Identificação de Gênero)

Automatically classifies the gender of proper names found in a textual dataset.

- **Library used:** `gender_guesser`
- **How to use:**
  1. Select the Identify Gender option.
  2. Upload a CSV file containing a list of names.
  3. ATAS generates a new CSV with the gender classification associated with each name.

4. Text Statistics (Brazilian Portuguese: Estatísticas de Texto)

Provides quantitative metrics such as **word frequency, named entities, and lexical diversity**. These metrics assist researchers in identifying thematic emphases, recurring actors, and stylistic features in academic texts.

- **Libraries used:** `spaCy`, `pandas`
- **How to use:**
  1. Go to the Text Statistics option.
  2. Select a text file.
  3. The system presents a detailed statistical report, including **word clouds and graphs**.

Note: Sentiment analysis is under development and will be integrated in future releases using models specifically trained for Brazilian Portuguese (e.g., Stanza, NLPNet, Udpipe).

5. Graphical Interface

ATAS offers an intuitive visual interface based on `tkinter` and `ttkbootstrap`, allowing users without programming knowledge to easily access its functionalities.

Implementation

ATAS is developed in **Python 3.8+** and utilizes:

- **`spaCy`** for text processing.
- **`pandas`** for data manipulation.
- **`tkinter`** and **`ttkbootstrap`** for the graphical interface.
- **`gender_guesser`** for gender identification.

While the current version relies primarily on `spaCy`, future developments will include support for alternative Natural Language Processing libraries (e.g., Stanza, NLPNet, Udpipe) to increase flexibility and expand coverage for Brazilian Portuguese. Sentiment analysis is also planned for future releases through the integration of models specifically trained for Portuguese. The source code is available on GitHub: https://github.com/AlidesChimin/SATA

References

Refer to the `paper.bib` file for the complete list of references.

Acknowledgments

I thank the project collaborators and the research groups GEPES and GETE, who influenced the conception of this software.

Zenodo DOI: 10.5281/zenodo.14868064

Bardin, L. (2011). *Análise de conteúdo*. Edições 70.

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An open source software for exploring and manipulating networks*. https://gephi.org/

Harvey, D. (1980). *A justiça social e a cidade*. Hucitec.

Harvey, D. (2005). *A produção capitalista do espaço*. Annablume.

Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016). *Who is doing computational social science? Trends in big data research*. SAGE Publishing. https://doi.org/10.4135/wp160926

Santos, M. (1978). *Por uma geografia nova: Da crítica da geografia a uma geografia crítica*. Hucitec.

Santos, M. (1996). *A natureza do espaço: Técnica e tempo, razão e emoção*. Hucitec.