The Journal of Open Source Software

ATAS - Academic Text Analysis System

Alides Baptista Chimin Junior [1]

1 Universidade Estadual do Centro-Oeste (UNICENTRO)

DOI: 10.21105/joss.08599

Software

- Review
- Repository
- Archive

Editor: Richard Littauer

Reviewers:

- @DiegoAscanio
- @rafaelanchieta
- @felipemaiapolo

Submitted: 25 February 2025
Published: 07 December 2025

Abstract

The ATAS – Academic Text Analysis System (Brazilian Portuguese SATA – Sistema de Análise de Textos Acadêmicos) is open-source software to support textual content analysis. Inspired by Bardin's method (Bardin, 2020), ATAS streamlines extraction, filtering, and statistical analysis of academic texts to surface semantic and linguistic patterns. It is particularly useful to Humanities and Social Sciences (Geography) researchers, offering tools for keywords, bigrams, lexical categories, and other quantitative text measures. ATAS integrates NLP workflows and exports data to tools such as Gephi for semantic network analysis (Bastian et al., 2009).

The toolset is localized for Brazilian Portuguese and is actively used by the GETE – Grupo de Estudos Territoriais (Territorial Studies Group) at UEPG and the GEPES – Grupo de Pesquisa Redes de Poder, Migrações e Dinâmicas Territoriais (Research Group on Power Networks, Migrations, and Territorial Dynamics) at UNICENTRO (Brazil), both operating in Master's and PhD programs.

Statement of Need

In Brazil's Humanities and Social Sciences (Geography), many scholars still rely on manual or semi-manual workflows to process large textual corpora. Historical barriers—limited programming experience, scarce infrastructure, and uneven training—have hindered the adoption of computational methods (Metzler et al., 2016). At the same time, Large Language Models (LLMs) have transformed NLP in recent years, opening powerful possibilities but also raising concerns about language coverage, transparency, and reproducibility (OpenAI, 2023; Scao et al., 2022). For Portuguese (Brazil) specifically, the ecosystem has improved with encoder models such as BERTimbau (Souza et al., 2020) and more recent PT-BR Encoders (Mello et al., 2024), yet accessible, GUI-driven tools that operationalize these advances for non-specialists remain scarce.

ATAS addresses this gap as an open-source, GUI-based system that automates key tasks—keyword filtering, bigram networks, lexical metrics—without requiring advanced coding. Localization to Brazilian Portuguese promotes equitable access to computational analysis and supports reproducible research pipelines for Human Geography and allied fields.

Beyond facilitating content analysis, ATAS helps investigate how discourse constitutes space. In Human Geography, space is socially produced and relational rather than a neutral container (Lefebvre, 1991; Massey, 2005). While GIS excels at mapping and spatial querying, its discrete data structures can under-represent processual and discursive spatialities. ATAS complements cartography by extracting semantic networks and entities directly from text, enabling researchers to track how places, actors, and relations are constructed, contested, and reconfigured in academic and policy 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` (Explosion AI, 2023)
- 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` (The Pandas Development Team, 2023)
- 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` (Hutchinson, 2016)
- 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 to surface 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 will be integrated in future releases with Portuguese-specific models (e.g., Stanza, NLPNet, UDPipe) and, where appropriate, open LLMs fine-tuned for PT-BR (Touvron et al., 2023), while preserving ATAS's principles of openness, transparency, and independence from proprietary APIs (OpenAI, 2023; Scao et al., 2022).

5. Graphical Interface

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

Installation

ATAS can be installed locally from source. The repository includes all dependencies in the `requirements.txt` file.

```
git clone https://github.com/AlidesChimin/SATA.git
cd SATA
pip install -r requirements.txt
python main.py
```

The software was tested under Python 3.8+ on Linux and Windows environments.

Reproducibility and Testing

All examples in the paper can be reproduced using the sample datasets available in the `/examples` directory. Each function (e.g., text filtering, bigram extraction, gender identification) can be validated independently. Automated tests are implemented through continuous integration on GitHub Actions, confirming successful imports, dependencies, and environment reproducibility.

Figure 1: Tests

Community Guidelines

The project follows open-source contribution standards. Users and contributors can:

1. Open issues and pull requests on GitHub.
2. Follow the guidelines described in the `CONTRIBUTING.md` file.
3. Adhere to the `CODE_OF_CONDUCT.md` included in the repository.

Scholarly Effort

ATAS comprises approximately 910 total lines of source code, of which 655 are Python logic, distributed across eight modules. The system includes 38 functions and 6 classes, integrating major open-source libraries (`spaCy`, `pandas`, `tkinter`, `ttkbootstrap`, `gender_guesser`, and `PIL`). It was developed by Dr. Alides Baptista Chimin Junior as part of ongoing research in the GEPES and GETE groups, within the field of Human Geography. The project reflects over a year of design, implementation, and integration efforts, bridging computational linguistics and content analysis for the Humanities.

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 add Stanza/NLP-Net/UDPipe backends and optional open LLM components for PT-BR tasks, keeping the stack reproducible and OSI-compliant. The source code is available on GitHub: https://github.com/AlidesChimin/SATA.

Acknowledgments

I thank the project collaborators and the research groups GETE (UEPG) and GEPES (UNICENTRO), whose ongoing use cases and feedback have shaped ATAS's development.

Figshare DOI: 10.6084/m9.figshare.30776753

References

Bardin, L. (2020). *Análise de conteúdo*. Edições 70.

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An open source software for exploring and manipulating networks*. https://doi.org/10.1609/icwsm.v3i1.13937

Explosion AI. (2023). *spaCy: Industrial-strength natural language processing in python* (Version 3.8). https://spacy.io/

Hutchinson, D. (2016). *Gender-guesser: Library for guessing the gender of first names*. https://pypi.org/project/gender-guesser/

Lefebvre, H. (1991). *The production of space*. Blackwell.

Massey, D. (2005). *For space*. SAGE.

Mello, J. P., Fonseca, L., & others. (2024). Advancing portuguese language models: The PT-BR encoders. *Journal of the Brazilian Computer Society*.

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

OpenAI. (2023). GPT-4 technical report. *arXiv Preprint*. https://arxiv.org/abs/2303.08774

Scao, T. L., Fan, A., & others. (2022). BLOOM: A 176B-parameter open-access multilingual language model. *arXiv Preprint*. https://arxiv.org/abs/2211.05100

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: Pretrained BERT models for brazilian portuguese. *arXiv Preprint*. https://arxiv.org/abs/2009.08144

The Pandas Development Team. (2023). *Pandas: Python data analysis library* (Version 2.3). https://pandas.pydata.org/

Touvron, H., Martin, L., Stone, K., & others. (2023). LLaMA 2: Open foundation and fine-tuned chat models. *arXiv Preprint*. https://arxiv.org/abs/2307.09288