A Multi-lingual Dataset of Classified Paragraphs from Open Access Scientific Publications

Eric Jeangirard[1]

[1]French Ministry of Higher Education and Research, Paris, France

October 2025

Keywords: open science, research data, research software, software mentions, data mentions, grobid, acknowledgments analysis, clinical trials mentions, scientific text mining

Abstract

We present a dataset of 833k paragraphs extracted from CC-BY licensed scientific publications, classified into four categories: acknowledgments, data mentions, software/code mentions, and clinical trial mentions. The paragraphs are primarily in English and French, with additional European languages represented. Each paragraph is annotated with language identification (using fastText) and scientific domain (from OpenAlex). This dataset, derived from the French Open Science Monitor corpus and processed using GROBID, enables training of text classification models and development of named entity recognition systems for scientific literature mining. The dataset is publicly available on HuggingFace https://doi.org/10.57967/hf/6679 under a CC-BY license.

Background & Summary

Scientific publications contain structured information that extends beyond the main research findings. Acknowledgments sections reveal funding sources and research infrastructures; data availability statements indicate research datasets; software mentions document computational tools; and clinical trial references provide links to registered studies. Automatically identifying and extracting these elements is crucial for research assessment, reproducibility studies, and understanding the research ecosystem.

While tools like GROBID can identify certain sections (e.g., acknowledgments) in English-language publications, coverage for other languages and paragraph types remains limited. This dataset addresses this gap by providing labeled paragraphs across multiple categories and languages, enabling the development and improvement of:

1. **Text classification models** to identify paragraph types in scientific publications, particularly for non-English languages
2. **Named entity recognition (NER) models** to extract:
   - Dataset names, DOIs, and accession numbers
   - Software names, repository URLs (GitHub, GitLab), and identifiers (SWHID, Wikidata)
   - Clinical trial identifiers (e.g., NCT numbers)
   - Funding agencies, grant IDs, infrastructures, and private entities in acknowledgments

The dataset draws from the French Open Science Monitor (FOSM), which tracks open science practices in France, and leverages existing text mining tools (GROBID, Softcite, DataStet) for semi-automatic annotation.

Methods

Source Corpus

The source publications were obtained from the French Open Science Monitor corpus, described in detail by (Bracco et al. 2022). The FOSM aggregates metadata and full-text content from French-affiliated scientific publications available under open licenses. Publications are primarily from the period 2013 onwards, ensuring contemporary research practices are represented.

Paragraph Extraction

Paragraphs were extracted from PDF documents using GROBID (GeneRation Of BIbliographic Data), a machine learning library for extracting and structuring raw documents into structured XML/TEI formats. GROBID's document segmentation capabilities identify distinct textual units within publications, including body paragraphs, acknowledgments, and supplementary sections.

Classification and Annotation

Paragraphs were semi-automatically classified into four categories using specialized tools:

1. **Acknowledgments**: Identified using GROBID's native section recognition, plus extra heuristics
2. **Data mentions**: Detected using DataStet, a tool for identifying dataset mentions in scientific text, plus extra heuristics
3. **Software/Code mentions**: Identified using Softcite, a tool for detecting software mentions, plus extra heuristics
4. **Clinical trial mentions**: Detected through pattern matching for trial identifiers

Softcite and Datastet are described in (Bassinet et al. 2023), they machine learning methods for entity recognition in scientific literature.

Note: No inter-annotator validation was performed. Users should be aware that classification accuracy depends on the performance of the underlying tools and may contain errors.

Language Identification

Language detection was performed using fastText's language identification model (lid.176.bin) (Joulin et al. 2016), which supports 176 languages. This pre-trained model provides efficient and accurate language prediction for short and medium-length texts. This detected language was kept instead of the language metadata provided by OpenAlex.

Scientific Domain Assignment

Each paragraph was assigned a scientific domain based on the OpenAlex classification system. Specifically, we used the `field` attribute from the `primary_topic` of the source publication. OpenAlex provides a hierarchical classification covering major scientific disciplines.

Data Structure

The dataset is distributed as a CSV file with the following columns:

- `license`: License, from OpenAlex data
- `text`: Text content of the paragraph
- `doi`: DOI of the publication
- `type`: Publication type, from OpenAlex data
- `detected_lang`: ISO 639-1 detected language code, with fasttext
- `publication_year`: Publication year

- `is_dataset`: Boolean, true if the paragraph mentions data
- `is_software`: Boolean, true if the paragraph mentions software or code
- `is_acknowledgement`: Boolean, true if this is an acknowledgement paragraph
- `is_clinicaltrial`: Boolean, true if the paragraphs mentions a clinical trial
- `field_name`: Field of the primary topic, from OpenAlex data
- `field_id`: OpenAlex id of the field

Data Records

The dataset is publicly available on HuggingFace at: https://doi.org/10.57967/hf/6679 The dataset contains 833k paragraphs distributed as follows:

- **Acknowledgments**: 108k paragraphs
- **Data mentions**: 570k paragraphs
- **Software mentions**: 203k paragraphs
- **Clinical trials**: 8.7k paragraphs

Language distribution:

- English: 98.4%
- French: 1.5%
- Other European languages: ~0.15%

License: CC-BY 4.0

Format: CSV (UTF-8 encoding)

Technical Validation

Classification Quality

As the dataset was produced through semi-automatic annotation without systematic manual validation, users should consider the following limitations:

1. **Tool-dependent accuracy**: Classification quality depends on the performance of GROBID, Softcite, and DataStet on the source documents
2. **Language bias**: Tools may perform better on English-language content than other languages
3. **Domain variability**: Performance may vary across scientific fields
4. **Boundary effects**: Paragraph segmentation may occasionally split or merge logical units

We recommend users conduct their own validation on a sample of the data relevant to their specific use case.

Language Identification Quality

FastText's lid.176.bin model has been shown to achieve high accuracy on well-formed texts. However, scientific texts may contain:

- Mixed-language content (e.g., English terms in French text)
- Technical terminology and symbols
- Short paragraphs with limited context

Usage Notes

Potential Applications

1. Paragraph Classification Models

The dataset can be used to train or fine-tune models for identifying paragraph types in scientific publications. This could be used to extend and improve upon GROBID's capabilities, particularly for non-English languages.

2. Named Entity Recognition

Paragraphs could help building training data for NER models targeting:

- **Data mentions**: Dataset names, DOIs, accession numbers
- **Software mentions**: Software names, repository URLs (GitHub, GitLab), identifiers (SWHID, Wikidata)
- **Clinical trials**: NCT identifiers and trial registries
- **Acknowledgments**: Funding agencies, grant IDs, research infrastructures, private entities

3. Research Ecosystem Analysis

The dataset enables large-scale studies of:

- Data sharing practices across disciplines
- Software usage patterns in research
- Funding acknowledgment practices
- Clinical trial documentation

Recommended Practices

1. **Data splitting**: Ensure publication-level splits (not paragraph-level) to avoid data leakage between train/test sets
2. **Multilingual evaluation**: Report performance separately for major language groups
3. **Domain-specific testing**: Validate models across different scientific fields
4. **Error analysis**: Manual inspection of misclassifications can reveal systematic issues

Limitations

- **No manual validation**: Classification errors from source tools are propagated
- **Imbalanced distribution**: Categories and languages may be unevenly represented
- **Temporal bias**: Predominantly recent publications (2013+) may not reflect historical practices
- **French affiliation bias**: FOSM focuses on French-affiliated research, which may not generalize globally
- **License restriction**: Only CC-BY content included; results may not generalize to closed-access literature

References

Bassinet, Aricia, Laetitia Bracco, Anne L'Hôte, Eric Jeangirard, Patrice Lopez, and Laurent Romary. 2023. "Large-scale Machine-Learning analysis of scientific PDF for monitoring the production and the openness of research data and software in France." https://hal.science/hal-04121339.

Bracco, Laetitia, Anne L'Hôte, Eric Jeangirard, and Didier Torny. 2022. "Extending the open monitoring of open science." https://hal.science/hal-03651518.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. "Bag of Tricks for Efficient Text Classification." *arXiv:1607.01759 [Cs]*, August. http://arxiv.org/abs/1607.01759.