RESEARCH REPORT:
RE3 and Iterator Datasets for Revision-Focused NLP Applications

INTRODUCTION

In recent years, there has been increasing interest in developing Natural Language Processing (NLP) systems that can understand and support iterative writing processes. This shift moves beyond static evaluation of text quality and aims to model how writers revise and improve their content over time. To support such advancements, high-quality revision datasets are critical. This research focuses on two publicly available and academically endorsed datasets: RE3 (Revision Extraction and Evaluation for Research Papers) and Iterator (Iterative Revision Dataset for Argumentative Writing). Both datasets capture different domains and types of revisions, offering complementary resources for the development of NLP models targeting revision detection, generation, and evaluation.

Dataset 1: RE3

Name: Revision Extraction and Evaluation for Research Papers (RE3)
Source: UKP Lab GitHub Repository
License: CC BY-NC-SA 4.0
Permission: Granted via direct communication with dataset owners (see screenshots in Appendix)
Description:
RE3 is a curated corpus comprising revision histories of academic research papers, collected from arXiv preprints. It provides pairs of paper versions (typically v1 and v2) and annotations that highlight sentence-level changes between them. The dataset captures fine-grained edits including addition, deletion, replacement, and reordering of text. These changes are aligned using a combination of sentence alignment algorithms and manual annotations, making the dataset suitable for edit classification, revision intention analysis, and academic writing improvement tools.
Features:

- Includes around **20,000 sentence pairs** from **academic texts**.
- Annotations label the **type of revision** and whether it **affects meaning** or clarity.
- Supports research in **argumentation mining**, **document comparison**, and **version control**.

Use Cases:

- Training NLP models to suggest or predict **academic writing edits**.
- Supporting **writing tutors** and feedback systems for scholarly authors.
- Analyzing how academic authors revise arguments and structure across drafts.

Dataset 2: Iterator

Name: Iterator – Iterative Revision Dataset for Argumentative Writing
Source: Vipul Raheja GitHub Repository
License: CC BY-NC 4.0
Permission: Granted via direct communication with dataset authors (see screenshots in Appendix)
Description:
Iterator is a large-scale dataset developed to model **student revision behavior** on argumentative essays. It contains **over 21,000 pairs of drafts and their revisions**, sourced from student submissions to Cambridge English's **Write & Improve** platform. Each entry includes the **original draft**, **revised version**, **system-generated feedback**, and in some cases, **human tutor feedback**. The dataset captures a diverse range of **linguistic improvements**, from surface-level grammar fixes to deeper argumentative restructuring.
Features:

- Focus on **iterative improvement** rather than one-time editing.
- Includes **automated and human feedback**.
- Covers a wide variety of topics and writing styles.
- High volume of **real student data**, enhancing model generalizability.

Use Cases:

- Developing **AI writing assistants** that provide meaningful revision feedback.
- Training models in **automated essay scoring** with revision tracking.
- Studying **student learning patterns** in language acquisition and academic writing.

Conclusion

Both RE3 and Iterator datasets contribute significantly to the field of revision-aware NLP. While RE3 emphasizes academic clarity and argument enhancement in scholarly texts, Iterator provides rich insights into how learners revise their essays based on feedback. Researchers and developers aiming to build tools for automated writing support, intelligent tutoring systems, or revision generation can benefit from integrating both datasets. With proper permissions obtained for academic usage, these datasets offer ethical, high-quality resources for advancing next-generation NLP applications focused on continuous writing improvement.

References

Raheja, V., et al. (2023). *Iterator: A Large-Scale Dataset for Iterative Revision of Argumentative Writing*.
Stab, C., et al. (2017). *Detecting Style and Scope of Revisions in Academic Writing*. Proceedings of the ACL.
UKP Lab RE3 GitHub Repository: https://github.com/UKPLab/re3
Iterator GitHub Repository: https://github.com/vipulraheja/iterater

Appendix: Permission Emails

External **Re: Permission Request to Use RE3 Dataset for University Project**

RQ

Ruan, Qian <ruan@ukp.tu-darmstadt.de>

Thursday, 24 July 2025 at 8:54 pm

To: YASHICA BASSI

Dear Yashica,

Thank you for your message.

You're very welcome to use the data. You can find detailed information in our paper [1], including links to download the dataset and access the code, which are available in the footnote on the first page (github: https://github.com/UKPLab/re3 data:https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/4300). You might also be interested in a follow-up paper [2], which introduces a larger dataset with edit annotations.

Both datasets are released under the CC BY-NC 4.0 license, so you're free to use them for research purposes. Kindly make sure to cite the relevant papers in any reports or publications.
Good luck with the project!

[1]. https://arxiv.org/abs/2406.00197
[2]. https://arxiv.org/abs/2410.02028

Best regards,
Qian

External **Re: Permission Request to Use Dataset for University Project**

ZM

Zae Myung Kim <zaemyung@gmail.com>

Saturday, 26 July 2025 at 9:41 am

To: YASHICA BASSI

To protect your privacy, some external images in this message were not downloaded. Download external images Go to Settings

Hi Yashica,

Yes, you can use the dataset for your research project. See the attached file. Please cite the papers in your work. Good luck with your project!

Best,

IteraTeR_plus.tar.gz

Zae

On Thu, Jul 24, 2025 at 6:36 PM YASHICA BASSI <s222622259@deakin.edu.au> wrote:

> Hi Zae,
>
> Thank you for your reply.
>
> I am referring to the **Iterator dataset** available in this repository: https://github.com/vipulraheja/iterater/tree/main/dataset.
>
> I would like to use this dataset for a university research project on revision chain analysis and data processing. Please let me know if there are any specific terms of use, license conditions, or acknowledgements we should include when using this dataset.
>
> Thank you very much for your help and guidance!
>
> Best regards,
> Yashica Bassi
> Deakin University