LREC-COLING 2024

CASIMIR: A Corpus of Scientific Articles enhanced with Multiple Author-Integrated Revisions

Léane Jourdan
Florian Boudin
Richard Dufour
Nicolas Hernandez

Contact: leane.jourdan@univ-nantes.fr

TALN
Nantes Université
LS2N

Context

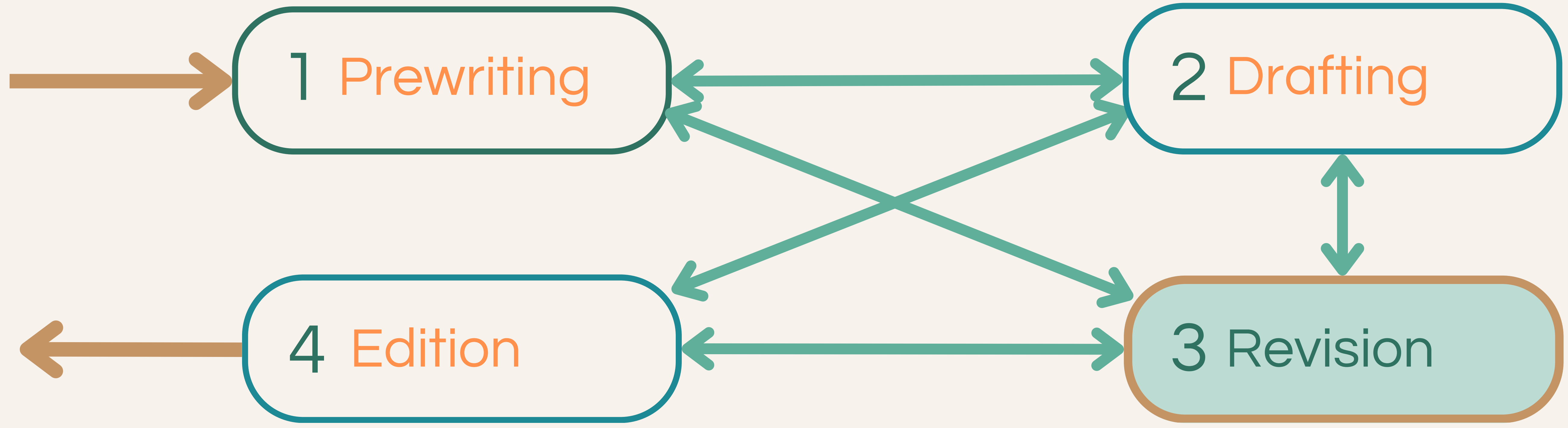
Motivations

- Writing an article is challenging
- Strong writing skills are essential
- Especially difficult for junior researchers and non-native English speakers

Domain

- Scientific writing assistance
- Focus on the revision step

The text revision task



Example:

The model has good results.

→

Our model shows good results in this task.

→

Our model shows excellent performance in this task.

(T.Ito et al, 2019 ; Seow, 2002 ; Laksmi, 2006 ; Bailey, 2014 ; Silveira et al., 2022 ; Du et al.,2022 ; Jourdan et al., 2023)

CASIMIR corpus

- **15 646** scientific articles with revisions
- **Alignment of the sentences and edits** between the versions of an article
- Enriched with article's **metadata and peer reviews**
- Exploitation for the training and evaluation of writing assistance tools

Example of revisions

Source text

*Recently, deep learning has **gained tremendous success** in modeling proteins, making data-driven **methods** more **appealing** than ever (Rives et al., 2019; Jumper et al., 2021).*
Nevertheless, challenges exist for** developing deep learning-based models to predict mutational effects on protein-protein **binding.
The major challenge is the scarcity of experimental data — only** a few **thousands of** protein **mutations** annotated with **the change** in binding **affinity** are publicly available (Geng et al., **2019b). This hinders** supervised learning **as** the **insufficiency of** training **data tends to cause over-fitting.

Revised text

*Recently, deep learning has **shown significant promise** in modeling proteins, making data-driven **approaches** more **attractive** than ever (Rives et al., 2019; Jumper et al., 2021).*
However,** developing deep learning-based models to predict mutational effects on protein-protein **binding is challenging due to the scarcity of experimental data.
Only** a few **thousand** protein **mutations,** annotated with **changes** in binding **affinity,** are publicly available (Geng et al., **2019b), making** supervised learning **challenging due to** the **potential for overfitting with insufficient** training **data.

Label of edits:

Content | **Language** | **Improve-grammar-Typo**

Comparison to existing corpora

	SMITH [1] 10/2019	IteraTeR [2] 03/2022	TETRA [3] 05/2022	F1000RD [4] 07/2022	arXivEdits [5] 10/2022	ARIES [6] 06/2023	CASIMIR 10/2023
Full-length articles				✓	✓	✓	✓
Contains articles with more than 2 versions		✓		✓	✓		✓
Real world revisions		✓		✓	✓	✓	✓
Peer reviews				✓		✓	✓
Large resource (> 2K revised articles)	?	✓					✓

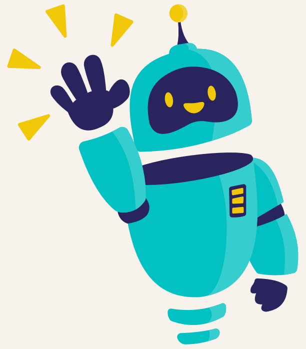
Table – Characteristics of previous datasets for scientific text revision compared to CASIMIR

[1] Ito et al. 2019; [2] Du et al. 2022; [3] Mita et al. 2022; [4] Kuznetsov et al. 2022; [5] Jiang et al. 2022; [6] D'Arcy et al. 2023

Summary

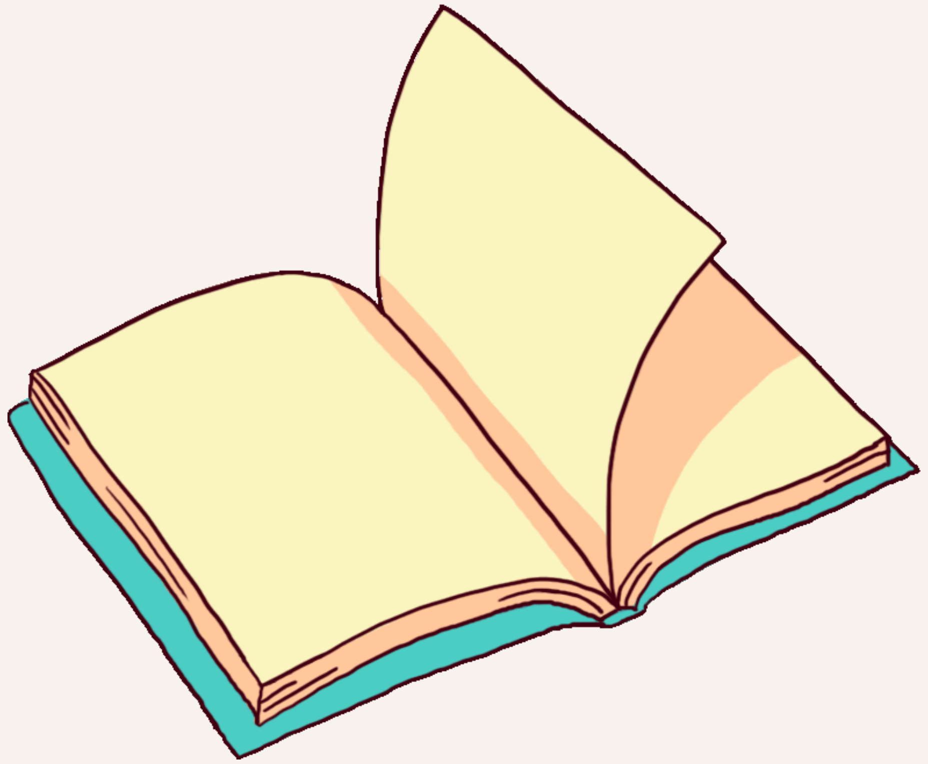
Corpus Creation

Qualitative Corpus Analysis



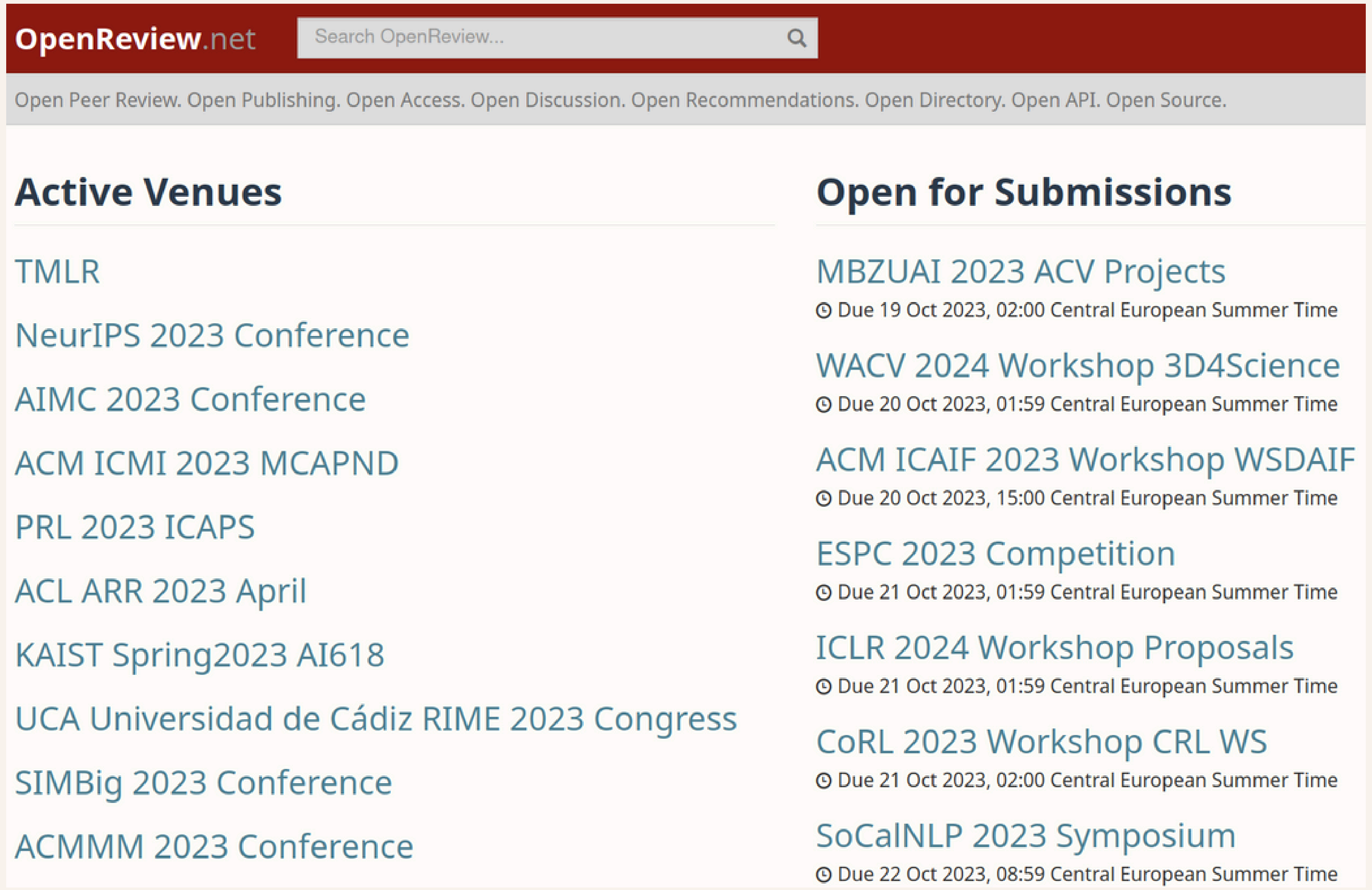
Experiments with Text Revision Models

1 – CREATION OF THE CASIMIR CORPUS

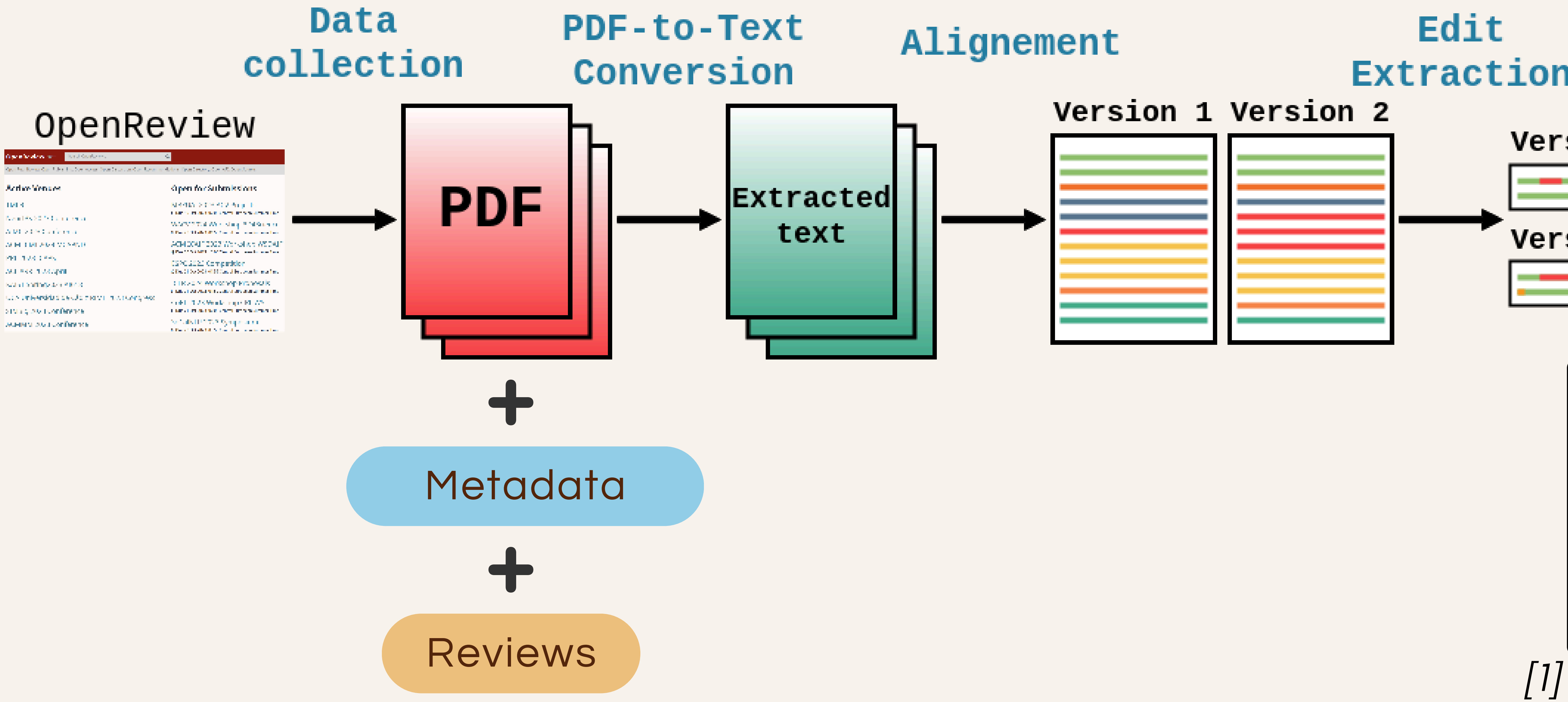


Creation of the casimir corpus

OpenReview



Creation of the casimir corpus

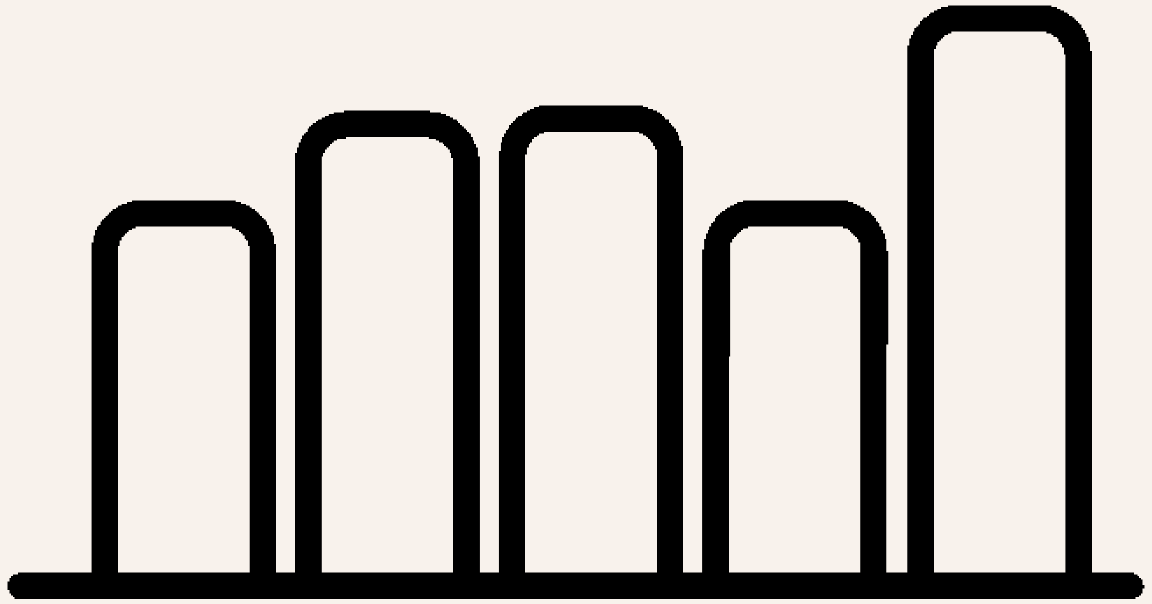


4 intentions possibles:
- Content
- Improve-grammar-Typo
- Format
- Language

[1] arXivEdits intention classifier

[1] Jiang et al. 2022

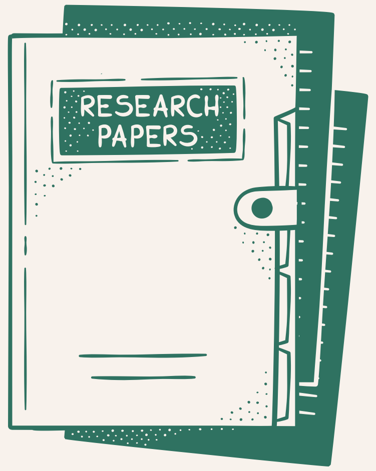
2 - CORPUS ANALYSIS



Content

Article pairs

- 15 646 different articles
- (3.5 versions by articles on average)
- 36 733 pairs of versions



CASIMIR

Metadata

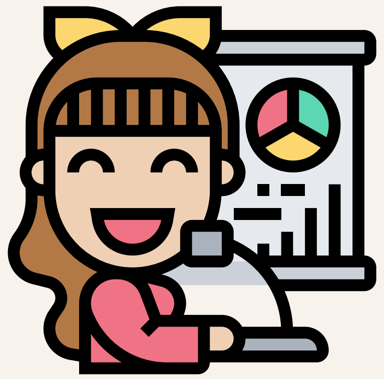
- Dates
- Authors
- Keywords
- Conference
- Ids...

29 conferences

Domains : machine learning (ICLR, ICML, NeurIPS), robotics (RSS, CoRL), NLP (ACL) and computer vision (ECCV)

Reviews

- Comments (can contain grades)
- Acceptance decision
- Dates...



Corpus analysis: Distribution of edits

5.2M of individual edit distributed in
3.7M of edited sentences

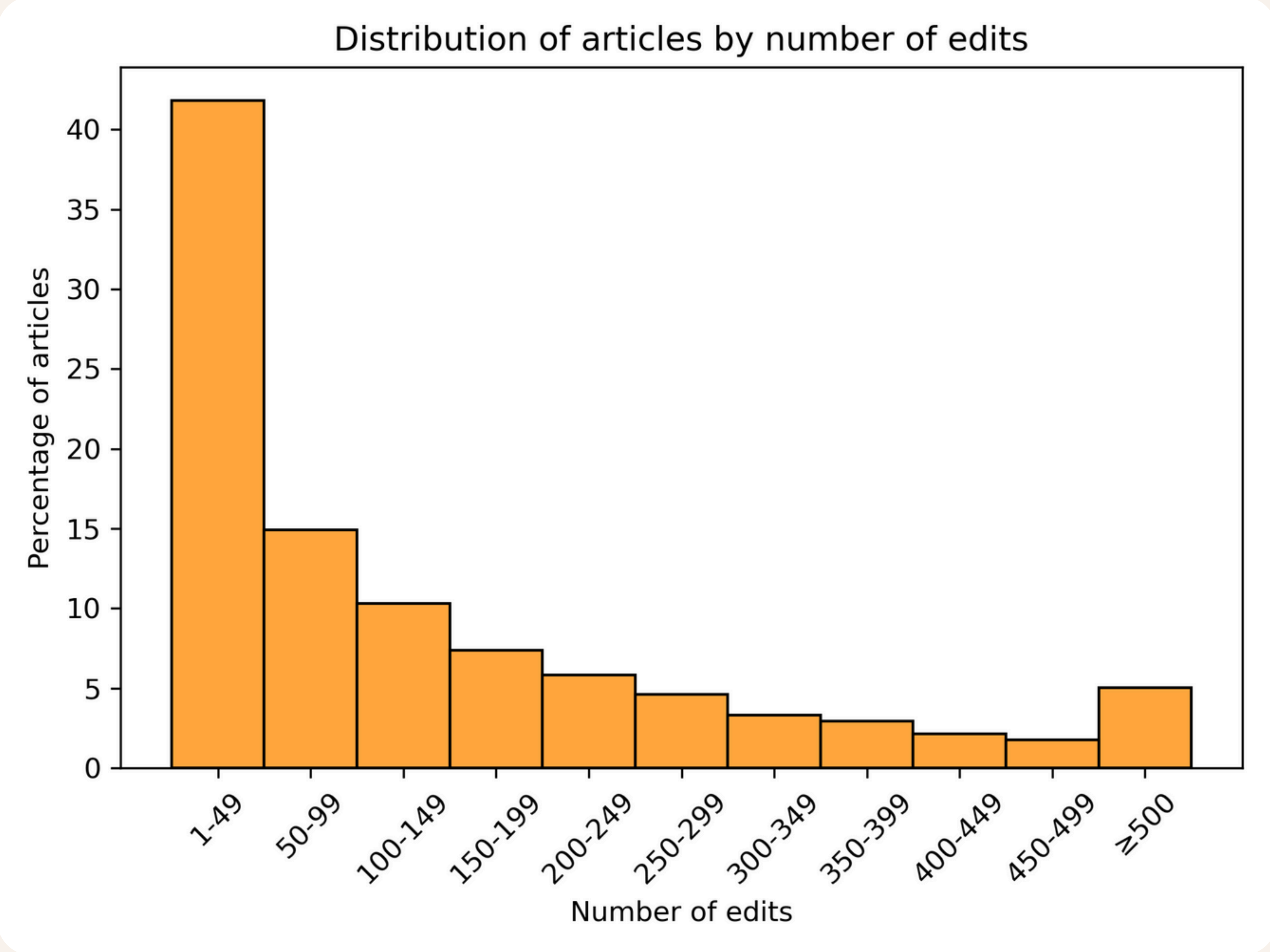


Figure 3: Distribution of articles by number of **edits**

Quantity of edits			
Min	1	First quartile	16
Max	4432	Median	74
Average	142.12	Third quartile	204
Edits length			
Min	1	Average	34.88
Max	9316	Median	13

Table 1: Distribution of the quantity of **edits by articles** and their length.

Edit intention	Percentage
Content	41.97%
Improve-grammar-typo	22.73%
Format	20.38%
Language	14.92%

Table 2: Distribution of **edit intentions**

Corpus analysis: Evolution and location of edits

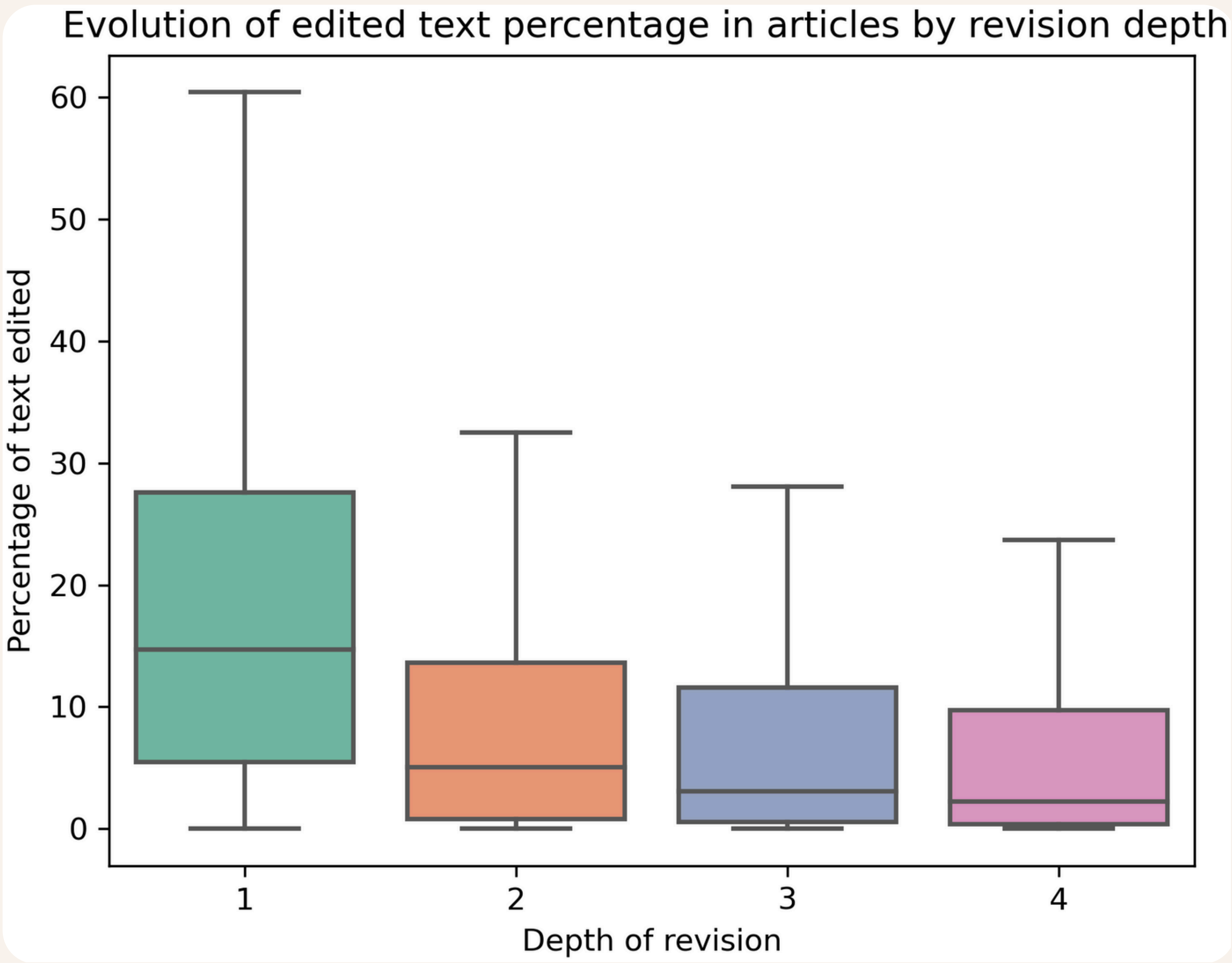


Figure 4: Evolution of edited text percentage in articles by revision depth.

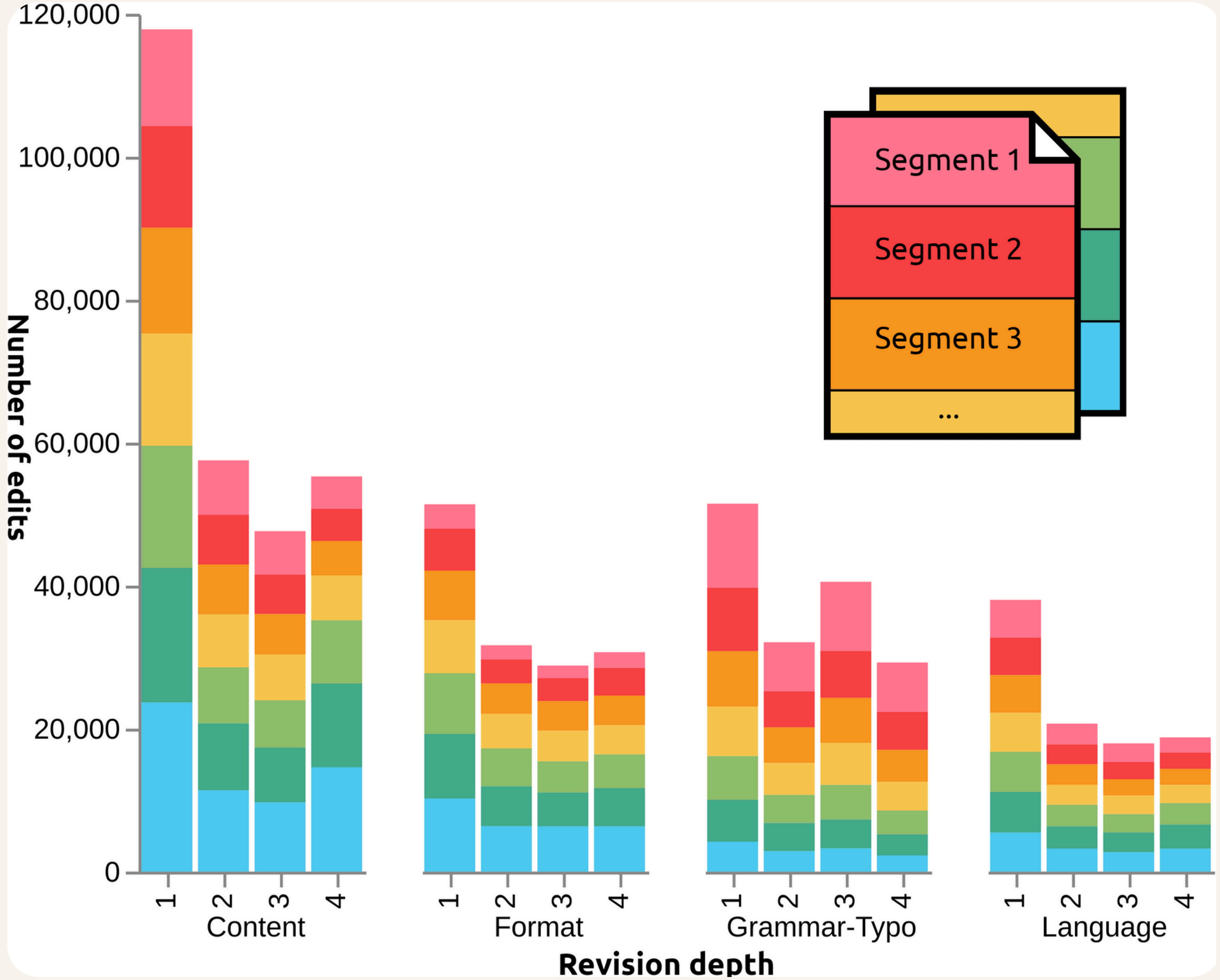


Figure 5: Evolution of the location of edited text by intention and revision depth

3 - EXPERIMENTS WITH TEXT REVISION MODELS

Experiments with Text Revision Models

Input:

A sentence to revised and an intention

Language

"To ***be able to study the performance of a*** *learned denoiser over a wide range of training set sizes we work with the ImageNet dataset (Russakovsky et al., 2015)."*

Output:

Generated revision

"To be able to study the performance of a learned denoiser over a wide range of training set sizes we ***use*** *the ImageNet dataset (Russakovsky et al., 2015)."*

The revised sentence

"To ***enable studying*** *learned denoiser over a wide range of training set sizes we work with the ImageNet dataset (Russakovsky et al., 2015)."*

Experiments with Text Revision Models

The tools

- IteraTeR-PEGASUS (Grammarly)
- CoEdIT (XL) (Grammarly)
- Llama2-7B (Meta)

The metrics

- Exact-match
- SARI
- BLEU
- ROUGE-L
- Bert-score

Every metric measure the similarity between the predicted sentence and the gold sentence.

(T.Ito et al, 2019; J.Li et al, 2022; Du et al.,2022)

Experiments with Text Revision Models

RESULTS

Model/Metric	EM	BLEU	ROUGE	SARI	BERT
CopyInput	0.00	**66.31**	**74.19**	61.38	94.46
Iterater-Pegasus (best intention)	6.04	60.99	73.25	55.27	95.93
Iterater-Pegasus (all intentions)	5.98	58.68	72.36	53.77	93.29
CoEdIT (best intention)	8.27	58.88	70.89	53.94	**96.08**
CoEdIT (all intentions)	8.25	56.44	69.22	51.62	95.99
Llama2-7B (best intention)♣	**14.05**	61.91	73.02	**62.07**	92.84
Llama2-7B (all intentions) ♣	13.76	57.46	68.18	58.39	92.37

Table 3: Results for all baselines. ♣ are results on the small set, others are realized on the large set.

Conculsion

LREC-COLING 2024

CASIMIR: A Corpus of Scientific Articles enhanced with Multiple Author-Integrated Revisions

Contact: leane.jourdan@univ-nantes.fr

TALN

taln.ls2n.fr

Article:

https://arxiv.org/abs/2403.00241

Corpus:

https://huggingface.co/datasets/taln-ls2n/CASIMIR