COCOA: Creation and Exploratory Investigation of a Corpus of Claims from NLP Articles

Clémentine Bleuze[1], **Fanny Ducel**[2], **Maxime Amblard**[1], **Karën Fort**[1]
[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France
{clementine.bleuze, maxime.amblard, karen.fort}@loria.fr, fanny.ducel@inria.fr

Abstract

Research articles are an essential pillar of scientific knowledge, but they are subject to multiple constraints. On the one hand, their scientific reliability is essential and relies in particular on the peer review process. On the other hand, they fulfill a rhetorical function of persuasion for authors who defend claims in a more and more competitive environment. In a context of massively increasing publication growth and quickly evolving practices, it is essential that the scientific community remains alert and critical of its own biases. In this paper, we call for a "NLP for NLP" framing of these issues. We created COCOA, a corpus of sentences from NLP papers and pre-prints published in English between 1952 and 2024, a sample of which we manually annotated with claim category labels reflecting their rhetorical function. We fine-tuned a SciBERT model to predict remaining labels, and made both the corpus and the model available to the community. We illustrate the interest of the corpus with exploratory analyses, and outline directions for further research. We hope that this work can stimulate discussions on the issues of research standardization and scientific overclaiming.

Keywords: argumentative zoning, claims, NLP4NLP, ethics

1. Introduction

Writing a research paper is never a neutral exercise. Although peer-reviewing is the standard to evaluate manuscripts from a scientific perspective, academic publishing is also an intrinsically rhetorical exercise implying that researchers emit and defend claims to an audience (Gilbert, 1976; Horton, 1995). This can become problematic when research results are presented more positively than justified by the obtained evidence, a phenomenon which has been referred to as "spin" in the medical literature (Koroleva, 2020). Indeed, similar to how "spin" has been shown to influence clinician's interpretation of studies' results (Boutron et al., 2014), exaggerated claims in NLP articles could propagate misleading representations of NLP systems (Inie et al., 2026) and their associated downstream impact (Reiter, 2025).

Besides, the increasingly popular tendency to share early pre-prints on social media (via open-access repositories such as arXiv[1]) challenges the traditional cycle of publication. Notably, Bagchi et al. (2025) found this practice to boost citation counts for authors, but they also stress that this could encourage the spreading of misinformation in unreviewed low-quality pre-prints. Therefore, analyzing the content and structure of research articles (or pre-prints) constitutes a relevant approach to understand the writing (or publishing) practices of a given field. As the NLP community is particularly well-equipped in methods to conduct "NLP for NLP" (Mariani et al., 2019a,b) analyses, we believe that such efforts are crucial to reflect on our own practices and reconsider our collective relationship to publication (Rawat and Meena, 2014).

With this in mind, we introduce COCOA, a corpus of English-written, sentence-segmented, full-text NLP works intended to serve the study of writing practices in NLP research over the last decades. COCOA also contains sentence-level annotations describing the type of claims written by researchers. Our contributions can be summarized as follows:

- We constitute COCOA, a corpus of 15M+ sentences extracted from the full-text of 87k articles and pre-prints published in the ACL Anthology[2] and arXiv between 1952 and 2024, freely distributed for researchers to reuse[3].
- We create a custom taxonomy of claim categories inspired by works in Argumentative Zoning, which we use to manually annotate more than 14k sentences from the corpus. We automatically predict the remaining labels by fine-tuning a SciBERT (Beltagy et al., 2019) model on this manual sample[4].
- We demonstrate the utility of COCOA by proposing exploratory analyses. We notably show that papers and pre-prints are increasingly discussing results in their abstract, and

[1] https://arxiv.org.

[2] https://aclanthology.org.

[3] see instructions for reuse in Section 8.

[4] *ibid*.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2387–2399
11–16 May 2026.

that abstracts tend to adopt stereotypical structures evolving over time.

- Finally, following a "NLP for NLP" perspective, we draft multiple research directions which can be explored using our corpus.

2. Related works

Rhetorics of scientific writing: There exists a strong academic interest in the question of 'proper' scientific writing, as shown by the plethora of guides explaining how to write a good academic title (Nair and Gibbert, 2016), introduction (Swales, 1981), specific paper section (Thompson, 1993), or more generally an entire research article (Patience et al., 2015; Labaree, 2024). At the level of discourse, Swales (1990)'s *CARS* model notably consists of three consecutive rhetorical *moves* to be used in paper introductions to defend the relevance of a given study: the establishment of a *territory* (an area of research), of a *niche* (a particular subject), and the *occupying* of that niche (the solution). At a finer-level, Horton (1995) argues that the persuasive power of an article can be influenced by the authors' tuning of multiple parameters such as the use of an active (vs. passive) voice, the choice and position of adverbs and adjectives, the choice of personal pronouns, self-citation, etc. The question of whether these strategies apply in a cross-cultural context has also been studied by Martín-Martín (2008), who analyzed the hedging devices used by English and Spanish-writing researchers to mitigate the force of their claims.

Argumentative Zoning: Argumentative Zoning (AZ) is an "analysis of the argumentative and rhetorical structure of a scientific paper" (Teufel et al., 2009). It can be modeled as a sequential sentence classification problem where each sentence is assigned a category reflecting its rhetorical purpose. Teufel et al. (1999) notably introduced a general annotation scheme for research papers comprised of seven such categories (e.g., BACKGROUND or AIM), which have alternatively been called "Discourse Role labels" by Lauscher et al. (2018). However, multiple schemes exist which all share similar concepts while expressing their own specificities, as has been surveyed by Schrader et al. (2023), which we took as a starting point for our own taxonomy. Existing works have also restricted the use of AZ to paper abstracts (Yamada et al., 2020; Cohan et al., 2019; Jin and Szolovits, 2018), which is particularly fitted for biomedical sciences publications where structured abstracts are prevalent.

Sources of NLP literature: The ACL Anthology is an essential source of NLP literature which spans a large number of domain conferences. Notably, the ACL OCL corpus (Rohatgi et al., 2023) contains metadata, PDF files, citation graphs and structured full-text of 73k ACL Anthology papers, which extends previous corpora such as the ARC (Bird et al., 2008) and AAN (Radev et al., 2009). In parallel, the open access repository arXiv is also becoming an increasingly important source of dissemination of NLP literature, despite being exempt from a standard peer-reviewing procedure. It allows researchers to publish pre-prints as well as published works, possibly in multiple versions which can be later edited. A regularly-updated corpus of arXiv papers is distributed by arXiv.org submitters (2024). In addition, a corpus of 1.9M full-text arXiv papers has been constituted by Saier et al. (2023) who suggest use cases such as citation recommendation and paragraph-level IMRaD[5] classification.

3. Corpus constitution

3.1. Papers and pre-prints collection

In order to span a broad range of NLP publications, we decided to use both the ACL Anthology and arXiv as data sources. We reused the ACL OCL corpus (Rohatgi et al., 2023) and collected the metadata and full-text of 71,286 papers published in the ACL Anthology between 1952 and 2022. For arXiv, we aimed at collecting pre-prints, that is, works that have not been published in peer-reviewed conferences or journals, which allowed us to consider pre-prints as a distinct genre in our corpus. We started by gathering the metadata of all arXiv papers from 1992-2024 which have been tagged under the Computational Linguistics category (*cs.CL*) from the freely accessible dataset of arXiv.org submitters (2024). We filtered out papers which have been assigned a journal or DOI (likely indicating that they have been published), and those which were already present in our ACL sub-corpus (based on title). We then queried the corresponding PDFs using `gsutil`[6] and converted them into XML files using GROBID (GRO, 2008–2024). At that stage, 3,382 documents could not be converted into XML files (e.g., due to insufficient visual quality leading to software errors).

3.2. Articles processing

Among all the obtained XML files, we further excluded posters (as they are not suited for structure analysis) and papers which were not written in English. The final corpus contains a total of

[5] Introduction, Methods, Results, Discussion.

[6] https://cloud.google.com/storage/docs/gsutil.

source	#papers (metadata)	#papers (XML files)	#papers (full-text extracted)	#sentences	sentences/paper (avg.)
ACL Anthology	71,286	71,286	58,456	9,339,173	159.76
arXiv	33,815	30,433	29,311	6,511,636	222.16
Total	105,101	101,719	87,767	15,850,809	180.60

Table 1: Statistics about COCOA.

15,850,809 sentences extracted from 87,767 papers (Table 1). We used the `en_core_web_sm`[7] model from the `spaCy` Python library to segment the text into sentences.

4. Corpus annotation

4.1. A taxonomy of claims in NLP papers

4.1.1. Why multiple claims categories?

We propose to define *claims* as statements emitted in research papers, which may deceive readers if presented in a misleading manner (e.g., if the performance of a system is said to be "excellent" when reported results are actually moderate). Importantly, we believe that claims do not solely consist of result statements (see discussion in Section 6.3). Instead, they can trigger a reader's expectations towards multiple aspects of the considered article (its contributions, its limitations, etc.), which we will refer to as claim *categories*. Our taxonomy of claim categories was designed to cover all sentences found in NLP papers' abstracts, introductions, result (and/or analysis, discussion) sections, and conclusions.

While we took inspiration from the work of Teufel et al. (1999, 2009) who applied Argumentative Zoning to Computational Linguistic papers, we encountered difficulties in applying these annotation schemes to the sentences of COCOA[8]. Thus, we decided to elaborate our own taxonomy.

4.1.2. Taxonomy constitution

We drew a random sample of papers from COCOA and confronted them with an initial "draft" taxonomy comprising five non-mutually exclusive claim categories. This initial proposal was then iteratively refined (by adding, deleting, or merging categories) during four manual annotation rounds involving a total of six annotators (2-4 per round) equipped with the Doccano (Nakayama et al., 2018) platform. The annotators consisted of two researchers, three PhD candidates and one Master student, all working in the domain of NLP. During each round, the annotators had to classify consecutive sentences from a selection of papers into one or more of the taxonomy's categories. Then, agreement was computed using pairwise Cohen's kappa (Cohen, 1960) and overall Krippendorff's alpha (Krippendorff, 2013), while discussions helped adjusting the categories names and definitions. We considered stability to be achieved at the fourth iteration, with a strong agreement $\kappa = \alpha = 0.81$ (Artstein and Poesio, 2008) between the two main annotators (cf. Table 6). The resulting final version of the taxonomy comprises seven claim categories, described in Table 2, as well as an additional NON-CLAIM tag. Our choice to discard purely "technical" statements as NON-CLAIM was motivated by the intuition that they are less likely to be exaggerated by authors, and by the desire to focus on "key" claims of a paper, which require less specialized knowledge to be examined under a critical eye.

4.2. Annotation of claims

4.2.1. Manual annotation

Two annotators who were involved in the elaboration of the final version of the taxonomy and reported to have strong agreement (Table 6) independently annotated 14,792 sentences from 158 articles drawn from the corpus[9]. This took an estimated 27 hours of work and resulted in the collection of 15,992 distinct annotations (595 sentences were labelled with at least two categories). We found that NON-CLAIM and RESULT were more prevalent, while IMPACT and DIRECTIONS were among the least-represented categories (Table 3).

[7] https://spacy.io/models/en#en_core_web_sm.

[8] It seems to us that the annotation scheme of Teufel et al. (1999) is more useful to describe the *source* of a given claim in a paper (e.g., OWN *vs.* OTHER) and its *stance* towards other claims (e.g., BASIS *vs.* CONTRAST), rather than the nature of its content. Although the updated version of Teufel et al. (2009) is closer to our efforts, it comprises many more tags and still misses categories to describe what we framed as IMPACT.

[9] This sample of 158 articles was crafted to cover equally both the ACL Anthology (52.7%) and arXiv (47.5%), as well as different time periods: <1994 (15.2%), 1994-2004 (29.7%), 2004-2014 (27.2%), >2014 (27.8%). This is to account for the imbalance in COCOA, which comprises a majority of recent, arXiv papers.

category	definition and examples from the corpus
CONTEXT	Claims providing context, background or explanations to the reader *'Whichever approach ends up being taken (as determined primarily by the writing system of the language in question), little attention is usually paid to pronunciation variants stemming from connected speech processes, hypoarticulation, and other phenomena typical for colloquial speech, mostly because the resource is seldom directly empirically derived.'* (Lukeš et al., 2018)
CONTRIBUTION	Claims depicting the nature of authors' contributions, objectives, the outcome of their work, and key elements of description of this outcome *'Our architecture is a variant of the Seq2seq model where two different decoders are used instead of only one of the original architecture.'* (Dinarelli and Grobol, 2019)
OUTLINE	Sentences used to draw the outline of a paper or to explain the content of a figure or table *'In Section 2, we will give an overview of the main advantages of this approach.'* (Och and Ney, 2001)
RESULT	Claims of results, either experimental or non-experimental; also analysis and discussion of these results, authors' opinions or important arguments *'We can see from this chart that the relative ranking of the models remain the same, except for sizes 1-3, where the probabilistic parser does better (or no worse than) the unlexicalized classifier-based models.'* (Eryiğit et al., 2008)
IMPACT	Claims of observed or anticipated impact of the presented work on people / on the scientific community *'We believe that this is a critical moment in the life of dialogue system research, and we anticipate exciting breakthroughs in the near future, leading to systems that are not only useful but also easy to use and accommodating, such that users will prefer them over alternative means of acquiring their information needs.'* (Glass and Seneff, 2003)
DIRECTIONS	Claims announcing future developments planned or suggested by the authors, possible continuations to the presented work *'As future work, we aim at investigating the impact of using additional linguistic information (such as part-of-speech tags) on LIHLA's performance.'* (Caseli et al., 2005)
LIMITATION	Claims of observed or anticipated limitations, flaws, drawbacks of some aspects of the presented work *'While the training process itself does not entail any additional memory or computation overhead compared to vanilla CLIP, the process of generating text rewrites using LLMs can be computationally expensive, requiring significant GPU resources and taking hours for large datasets.'* (Fan et al., 2023)
NON-CLAIM	Any other sentence of a paper, including methodology, technical details, etc. *'We randomly selected a dataset of 150 tweets which were annotated by both annotators for both POS tagging and dependency structures.'* (Bhat et al., 2018)

Table 2: Final version of our taxonomy of claim categories, with examples from COCOA.

4.2.2. Automatic annotation

We used our manually annotated corpus as a training set for the automatic identification of claims, seen as a sentence-level, multi-label classification problem with eight classes. We led experiments to test the capabilities of traditional Machine Learning algorithms[10] (Logistic Regression and Support Vector Machines) with multiple vectorization techniques and input formats, but achieved limited performance[11]. We then considered three Pretrained Language Models (PLM) based on BERT (Devlin et al., 2019), commonly used in related literature: RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021) and SciBERT (Beltagy et al., 2019).

We fine-tuned these models from their HuggingFace endpoints[12] using a 8-1-1 ratio for train/dev/test splitting. We trained each model for 15 epochs with an early stopping strategy to prevent overfitting. As an input to the model, we compared three formats: (i) the sole target sentence, (ii) the target sentence with its preceding and succeeding

[10]We used scikit-learn for implementation.

[11]We found that a SVM model (with sigmoid kernel and regularization set to 5) with *bag-of-words* tokenization achieved the highest average weighted F1-score of 0.68.

[12]We used FacebookAI/roberta-base, microsoft/deberta-v3-base and allenai/scibert_scivocab_uncased.

category	count	rel. %
CONTEXT	2,254	14.1%
CONTRIBUTION	2,032	12.7%
OUTLINE	376	2.4%
RESULT	3,572	22.3%
IMPACT	154	1.0%
DIRECTIONS	453	2.8%
LIMITATION	555	3.5%
NON-CLAIM	6,596	41.2%
total	15,992	100%

Table 3: Category-wise counts of manual annotations.

category	F1-score
CONTEXT	0.93
CONTRIBUTION	0.87
OUTLINE	0.88
RESULT	0.86
IMPACT	0.52
DIRECTIONS	0.82
LIMITATION	0.51
NON-CLAIM	0.95
avg (weighted)	**0.89**

Table 4: Average and category-wise classification results of the best fine-tuned SciBERT model.

sentences, (iii) the target sentence with the two preceding sentences. We compared the models in terms of weighted F1-score and found that SciBERT obtained the best global performance (0.89) when used with input format (iii) the target sentence with the two preceding sentences. Despite this high result, we note that less-populated categories IMPACT and LIMITATION suffer from poorer scores (resp. 0.52 and 0.51, cf. Table 4).

We used this fine-tuned SciBERT model for inference on the rest of the corpus to obtain silver labels. In order to assess the quality of the predicted labels, we drew a sample of 100 articles and manually reviewed the predictions assigned to their abstract sentences[13]. Out of 604 sentences, we identified 11 outliers originating from XLM parsing errors (e.g., footnote elements identified as abstract sentences). For the remaining 593 sentences, the first author obtained 87.35% of strictly equal annotations, which represents an agreement of $\kappa = \alpha = 0.83$ with the model. Disagreement mainly came from a single paper where a sequence of CONTEXT claims was erroneously labeled as CONTRIBUTION by the model, and from cases of confusion between the double label CONTRIBUTION+RESULT and single labels CONTRIBUTION and RESULT. This last issue was frequently encountered during the manual annotation rounds. We believe this demonstrates the complexity of this qualitative task, while still providing us with a reasonable guarantee that silver labels are reliable enough for further processing.

[13] It would admittedly have been more reliable to check entire articles. However, using abstracts allowed us to augment the diversity of manually verified articles.

5. Corpus analysis

After running the model on the sentences of the corpus which had not been manually annotated, we find that 302,213 sentences have been assigned at least two labels, which leaves us with a total of 16,126,896 annotations on 15,850,809 sentences.

5.1. Categories characteristics

Distribution of categories: At corpus-level, we find that predicted categories follow a similar distribution as that of the manually annotated subcorpus (Table 5), although NON-CLAIM labels are relatively more represented. If we consider occurrences in papers (i.e., whether a category appears at least once), we find that more than 98% of papers contain at least one claim from the categories NON-CLAIM, CONTEXT, CONTRIBUTION and RESULT. More than 71% contain LIMITATION and DIRECTIONS, while only 47% (resp. 49%) contain OUTLINE (resp. IMPACT) claims.

category	count	rel. %
CONTEXT	1,440,275	8.9%
CONTRIBUTION	1,107,615	6.9%
OUTLINE	142,607	0.9%
RESULT	2,530,290	15.7%
IMPACT	112,305	0.7%
DIRECTIONS	261,512	1.6%
LIMITATION	273,614	1.7%
NON-CLAIM	10,258,678	63.6%
total	16,126,896	100%

Table 5: Category-wise counts of all annotations in the corpus.

Pairwise cooccurrences: Among sentences tagged with multiple labels, we find that the most frequent pairs are LIMITATION+RESULT (39% of all LIMITATION occurrences), IMPACT+DIRECTIONS (14% of all IMPACT occurrences) and OUTLINE+CONTRIBUTION (13% of

CONTEXT	Discourse relations bind smaller linguistic units into coherent texts.
CONTEXT	Automatically identifying discourse relations is difficult, because it requires understanding the semantics of the linked arguments.
CONTEXT	A more subtle challenge is that it is not enough to represent the meaning of each argument of a discourse relation, because the relation may depend on links between lower-level components, such as entity mentions.
CONTRIBUTION	Our solution computes distributed meaning representations for each discourse argument by composition up the syntactic parse tree.
CONTRIBUTION	We also perform a downward compositional pass to capture the meaning of coreferent entity mentions.
RESULT	Implicit discourse relations are then predicted from these two representations, obtaining substantial improvements on the Penn Discourse Treebank.

Figure 1: Example of an abstract from the corpus (Ji and Eisenstein, 2015) with a structure formed of three sequences: CONTEXT-CONTRIBUTION-RESULT. The annotations were generated automatically.

all OUTLINE occurrences). We believe that this reflects frequent phenomena, e.g., respectively: the discussion of limitations arising from the results of an experiment, an anticipation of a study's future impact if pursued in future work, and the simultaneous declaration of a paper's plan and of the main contributions presented in each part.

Linguistic characteristics: We analyze the average length, the vocabulary (most frequent terms), and named entities in each category[14]. We find certain terms to be quite specific to certain categories, like "section" (OUTLINE), "future" (DIRECTIONS), "et" and "al" (CONTEXT), although most frequent terms largely overlap between categories. We note that IMPACT claims are the longest (169 characters on average) and OUTLINE are the shortest (100 characters). In addition, CARDINAL entities are more present in CONTEXT, CONTRIBUTION and RESULT claims, which can be hypothesized to correspond with reports of quantitative results, corpus sizes, etc. Besides, DATE and PERSON are quasi-exclusive to CONTEXT claims, which likely corresponds to citations displaying authors' names and publication years.

5.2. Abstracts' structure

As a practical use case of our corpus to study NLP papers, we propose to analyze the structure of abstracts through time. We define the structure of an abstract as an ordered chain of sequences, that is, consecutive sentences sharing the same claim category (see Figure 1). We believe abstracts to be particularly interesting as they showcase a paper's core claims, but this could naturally be extended to other sections such as introductions or conclusions.

[14]We use the model https://spacy.io/models/en#en_core_web_sm. For a list of named entities supported by spaCy, with definitions, see https://github.com/explosion/spaCy/discussions/9147.

What's (not) in an abstract? We first notice that some categories tend to be absent from papers' abstracts: less than 9%, 4% and 3% of them contain IMPACT, LIMITATION and DIRECTIONS claims. They are however respectively present in the full-text of more than 45%, 70% and 78% of papers. This confirms the particular role of abstracts to highlight certain claim categories over others, which are preferably stated in results, analysis, discussion and concluding sections.

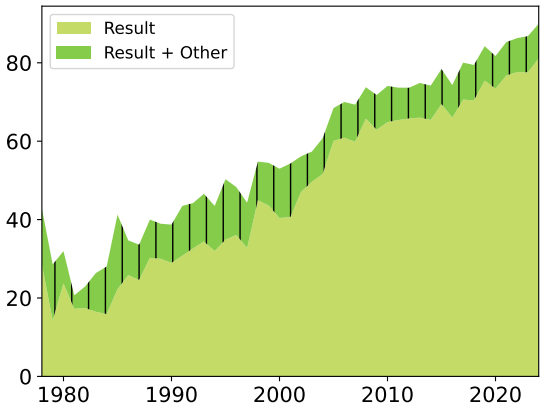

Figure 2: Share of papers (%) with at least one RESULT sequence in their abstract, per year of publication. We also consider multi-labels comprising RESULT. For readability, we start the plot in 1978, as previous years comprise less than 10 papers.

Reports of results are on the rise: We observe a drastic paradigm change as to the presence of RESULT claims in abstracts. Whereas only 42.86% of papers published in 1978 contained at least one RESULT sequence, this proportion has reached 89.95% in 2024. The trend has increased steadily, as can be visualized in Figure 2. Importantly, we do not observe such marked changes for other claim categories.

Longer, more standardized abstracts: We propose in Figure 3 a visual representation of the distribution of abstract structures among three consecutive periods of the corpus (1952-1994, 1995-2009 and 2010-2024). We note a progressive standardization of writing practices with reduced options for sequences 1-3 over time. Additionally, CONTEXT sequences progressively gain authors' preference over CONTRIBUTION sequences in the first position. We can observe the increasing place taken by RESULT sequences in abstracts, as was discussed in the previous point, as well as an intensified use of alternations between CONTRIBUTION and RESULT, sometimes in the form of double-sequences CONTRIBUTION+RESULT. Finally, abstract structures progressively lengthen: the most frequent structure before 1995 consists of a single CONTRIBUTION sequence, which becomes CONTRIBUTION-RESULT between 1995-2009, and extends to CONTEXT-CONTRIBUTION-RESULT between 2010-2024. Interestingly, we note a possible mapping between this last structure and the IMRaD structure commonly observed in health sciences since the 1970s (Sollaci and Pereira, 2004). It is indeed standard in this domain to produce well-structured papers and abstracts, which have been shown to improve readability and clarity (Sharma and Harrison, 2006). Although NLP conferences allow authors to write free form abstracts, this alignment to other disciplines' practices may be related to the diversification of NLP subfields and a desire for authors to make their work more easily understandable to a wider audience.

6. Research directions

In what follows, we discuss several research directions which we believe our corpus can help explore.

6.1. Article rhetorical structure analysis

As we have shown in Section 5.2, it is easy to leverage claim category labels to identify temporal trends in the structure of NLP papers. We believe that our preliminary study of abstracts can be deepened and generalized to other paper sections, depending on researchers' interests. If mapped with additional metadata such as topic keywords, our corpus could also allow to study structures among specific NLP subfields. For instance, researchers could study the structure of conclusions in papers related to Machine Translation.

It would be interesting to articulate this approach with other studies related to the domains of Rhetorics and Argumentation Mining (Lawrence and Reed, 2019). Indeed, some researchers have worked within sentences and investigated syntactical moves as relevant stylistic devices (Al Khatib et al., 2020). Others have complemented Discourse Role labels similar to those of our taxonomy with additional annotation layers for scientific texts, notably describing Citation Context and Subjective Aspect (Lauscher et al., 2018).

6.2. Differences between ACL articles and arXiv pre-prints

Until now, we have not extensively taken the origin of papers (ACL Anthology vs. arXiv) into consideration in our analyses. It is yet possible that these sub-corpora exhibit stylistic differences, even more so as some arXiv pre-prints may have been rejected by journals or conferences. We nonetheless make two preliminary observations:

- Abstract structures are very similar in both sub-corpora between 2010-2022[15], despite a stronger preference in arXiv papers for structures starting with CONTEXT rather than CONTRIBUTION.
- arXiv pre-prints are significantly more likely to contain sequences of IMPACT (55.18%, vs. 45.96%) and LIMITATION (75.30% vs. 70.25%), but less likely to contain OUTLINE sequences (42.63% vs. 49.14%) than ACL Anthology papers. This could be explained by a more formal style in published works, but requires further analysis.

We believe it would be valuable to study the ACL Anthology and arXiv sub-corpora independently to try and identify potential structural and/or linguistic differences more systematically. This is particularly important as arXiv benefits from a wide exposure, which may surpass that of conference proceedings in the long run (Bagchi et al., 2025).

6.3. Overclaiming detection

Overclaiming or "spin" happens when excessive enthusiasm leads authors to formulate inaccurate reports, interpretations, or extrapolations with regards to actual research results (Koroleva, 2017). Methods have been developed to identify spin in medical research (Koroleva, 2020), however they are not suited for unstructured, less-standardized NLP papers. In what follows, we call for a NLP-compatible investigation of overclaiming. We argue that categories from our taxonomy[16] can be asso-

[15] We choose this period as the histograms of both sub-corpora follow comparable distributions. Before 2010, we lack of arXiv pre-prints, and after 2022, of ACL Anthology papers.

[16] Excepted for NON-CLAIM and OUTLINE. We nevertheless included this last category as it is necessary to describe fully paper abstracts and introductions.

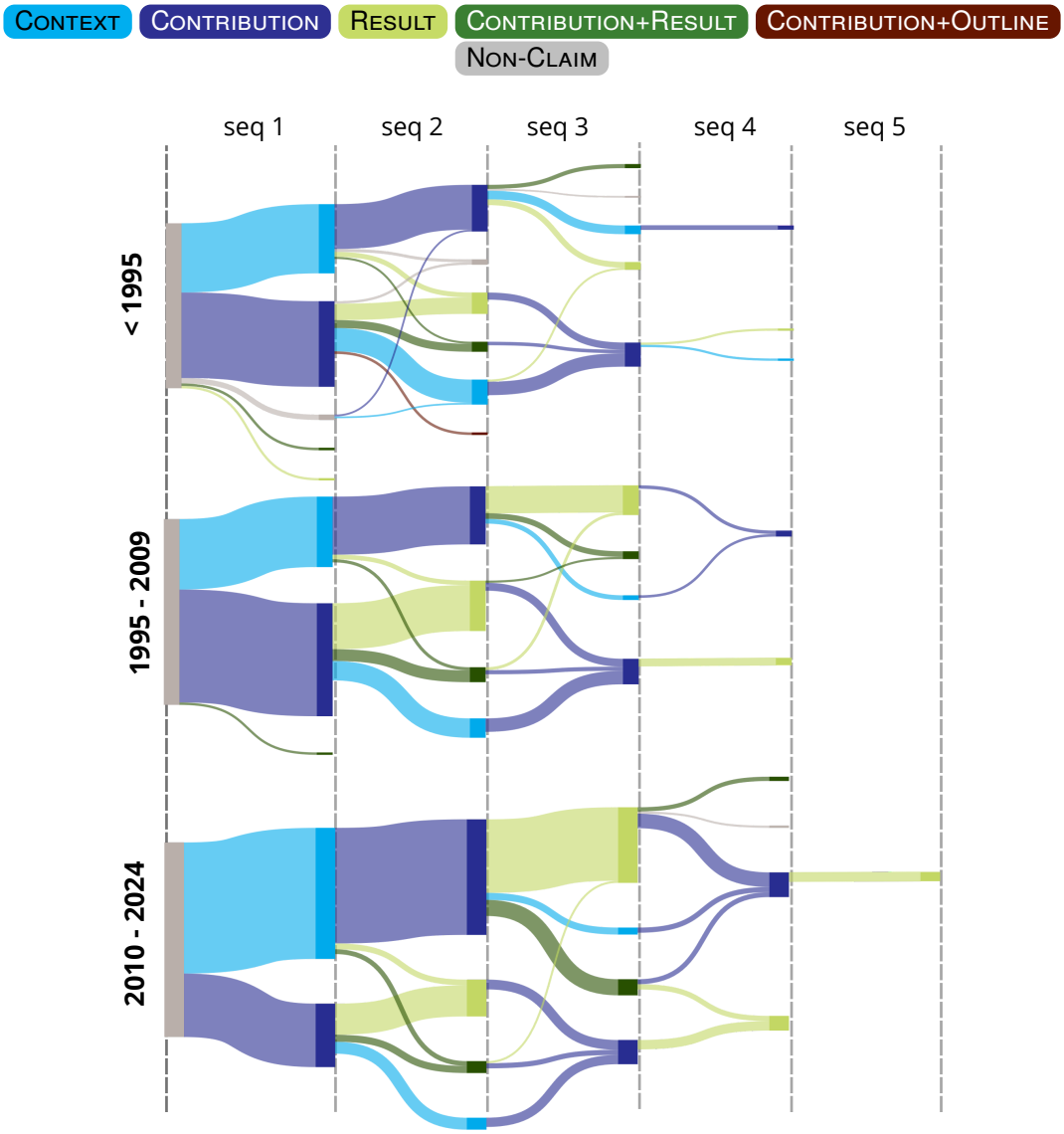

Figure 3: Sankey diagrams depicting the distribution of abstract structures among time periods in the corpus. For a chosen time period, diagrams should be read from left to right like trees representing the step-by-step construction of an abstract (see color legend above). Flow width is proportional to the share of papers adopting a given sequence (noted seq.) for the considered time period. For instance, papers published between 2010-2024 tend to start with a CONTEXT sequence (67.4% of cases), represented by a cyan flow, while the remaining 32.6% start with a CONTRIBUTION sequence, represented by a darker blue flow. For readability, we did not represent flows representing less than 1% of a node's size.

ciated with diverse forms of overclaiming that are not limited to RESULT claims.

For instance, Jouitteau and Grobol (2024) refuted the alleged ability of translation model m2m100 to "translate directly between any pair of 100 languages" (Fan et al., 2021), which in that case constitutes a CONTRIBUTION overclaim. They indeed observed that m2m100 was unable to translate between French and Breton, whereas these two languages were announced to be covered. According to Jouitteau and Grobol (2024), this case illustrates a *diversity washing* strategy commonly applied to languages with an intermediate numerical coverage such as Breton. We also consider that Rosenberg et al. (2023) made an exaggerated IMPACT claim that their system "provide[s] a possible pathway towards building a general-purpose Collective Superintelligence", when in reality, it only succeeded in a single, limited task: that of predicting how many gumballs are contained in a jar, based on a photograph.

Following these considerations, we propose that an automated system could rely on clues to identify papers at risk of overclaiming:

- the number, absence or presence of certain categories (e.g., papers with no LIMITATION claims but numerous IMPACT and DIRECTIONS claims)
- intra-sentence clues for "core" claims (e.g., assess whether RESULT claims in conclusion contain precise numbers vs. hedging clues showing uncertainty)
- consistency checks between key sections of papers (e.g., map and compare all CONTRIBUTION and RESULT claims from the abstract, introduction and conclusion)

Admittedly, it is difficult (and probably not desirable) to set objective and consensual criteria to define overclaiming. However, we hope to foster discussions as to how research is written and promoted by authors, including ourselves.

7. Conclusion

We presented COCOA, a freely available corpus of 15M+ sentences extracted from 87k NLP articles and pre-prints published between 1952 and 2024. COCOA contains claim category annotations from a custom taxonomy designed to suit NLP papers. A sample of 14,792 sentences has been manually annotated in order to fine-tune a SciBERT (Beltagy et al., 2019) model, which predicted the remaining labels. We presented exploratory analyses on the corpus and proposed a diachronic study of the structure of abstracts, which revealed an increasing focus on results. Finally, we suggested that article structure analysis, systematic stylistic comparison of ACL Anthology papers vs. arXiv pre-prints, and overclaiming detection are important research directions to be considered under a "NLP for NLP" perspective.

8. Reusing COCOA and our fine-tuned SciBERT model

Both COCOA[17] and our fine-tuned SciBERT model for claim category classification[18] are available for reuse under a CC BY-NC 4.0 license.

9. Ethical considerations and limitations

Ethical considerations: This work was motivated by a desire to account for rhetorical phenomena within NLP research articles. This is part of a larger context which raises issues related to the ever-increasing number of publications, the risk of spreading false information or lower-quality work, and the directions in which certain policies may lead public research. We believe that the NLP community has a duty to identify its own biases and potential for harm, and try to initiate collective discussion as to how to reduce them, aside from the pressure of current trends. Notably, Large Language Models (LLMs) constitute an eloquent example of widely promoted tools, which are now adopted way beyond the NLP field that developed them, despite reported issues of inaccuracy, lack of reproducibility, biased outputs, and elevated environmental impact. For these reasons, we did not attempt to use generative LLMs to label our corpus, and found a SciBERT model sufficient for our objectives.

Limitations: Despite our best efforts, COCOA is not exempt from imperfections. First, it does not constitute an exhaustive collection of NLP papers and pre-prints, due to missing documents, errors in PDF to XML conversions, and existing literature outside the ACL Anthology and arXiv. In addition, the latest years are not represented in COCOA, and we would like to be able to provide future updated versions spanning 2022-2026. We also acknowledge that the labels provided (both manual and automatic) can be subject to discussion due to model errors and intrinsic task subjectivity. Finally, the analyses we presented in Section 5 could be extended and further refined. In particular, we could use relative frequencies instead of raw frequencies to better account for "keyness" of certain words towards certain claim categories. However, we still believe that the present work constitutes a substantial contribution for the community, which we promptly invite to get a handle on, extend, criticize, or improve it according to their needs.

10. Acknowledgements

This work has received support from the French National Research Agency (Agence Nationale de la Recherche) under grant agreements ANR-20-SFRI-0009 (ORION) and ANR-23-IAS1-0004 (InExtenso). We also benefited from an access to the computing infrastructure Grid5000[19]. We would like to thank the annotators who participated in elaborating the taxonomy along with paper authors, namely Amandine Decker (CLASP, University of Göteborg, Sweden and LORIA, University of Lorraine, France) and Valentin Richard (LORIA, University of Lorraine, France, and ILLC, University of Amsterdam, the Netherlands). Finally, we thank the reviewers for their useful comments.

[17] https://doi.org/10.57967/hf/7946.
[18] https://doi.org/10.57967/hf/4797.
[19] https://www.grid5000.fr.

11. Bibliographical References

2008–2024. Grobid. https://github.com/kermitt2/grobid.

Khalid Al Khatib, Viorel Morari, and Benno Stein. 2020. Style analysis of argumentative texts by mining rhetorical devices. In *Proceedings of the 7th Workshop on Argument Mining*, page 106–116, Barcelona, Spain. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Chhandak Bagchi, Eric Malmi, and Przemyslaw Grabowicz. 2025. Effects of research paper promotion via arxiv and x. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 160–177, Copenhagen, Denmark. AAAI Press.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 987–998, New Orleans, Louisiana. Association for Computational Linguistics.

Isabelle Boutron, Douglas G. Altman, Sally Hopewell, Francisco Vera-Badillo, Ian Tannock, and Philippe Ravaud. 2014. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the spiin randomized controlled trial. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 32(36):4120–4126.

Helena M. Caseli, Maria G. V. Nunes, and Mikel L. Forcada. 2005. LIHLA: Shared task system description. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 111–114, Ann Arbor, Michigan. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Dinarelli and Loïc Grobol. 2019. Seq2Biseq: Bidirectional Output-wise Recurrent Neural Networks for Sequence Modelling. In *CICLing 2019 - 20th International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France.

Gülşen Eryiğit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. In *Advances in Neural Information Processing Systems*, volume 36, page 35544–35575, New Orleans, Louisiana. Curran Associates, Inc.

G Nigel Gilbert. 1976. The transformation of research findings into scientific knowledge. *Social studies of science*, 6(3-4):281–306.

James Glass and Stephanie Seneff. 2003. Flexible and personalizable mixed-initiative dialogue systems. In *Proceedings of the HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing*, pages 19–21, Edmonton, Canada.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, Online.

Richard Horton. 1995. The rhetoric of research. *BMJ*, 310(6985):985–987.

Nanna Inie, Peter Zukerman, and Emily M. Bender. 2026. De-anthropomorphizing "ai": From wishful mnemonics to accurate nomenclature. *First Monday*.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 3100–3109, Brussels, Belgium. Association for Computational Linguistics.

Mélanie Jouitteau and Loïc Grobol. 2024. Petits oublis, grands effets : le silençage des communauté linguistiques minorisées dans le tal et ses conséquences. In *Journée d'étude Journée Ethique et TAL 2024*, Nancy, France.

Anna Koroleva. 2017. Vers la détection automatique des affirmations inappropriées dans les articles scientifiques (towards automatic detection of inadequate claims in scientific articles). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, page 135–148, Orléans, France. ATALA.

Anna Koroleva. 2020. *Assisted authoring for avoiding inadequate claims in scientific reporting*. phdthesis, Paris-Saclay University; University of Amsterdam.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*, 3rd edition edition. Sage, Thousand Oaks, CA.

Robert V. Labaree. 2024. Research guides: Organizing your social sciences research paper. Url: https://libguides.usc.edu/writingguide/.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018. Arguminsci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, page 22–28, Brussels, Belgium. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv:1907.11692 [cs].

David Lukeš, Marie Kopřivová, Zuzana Komrsková, and Petra Poukarová. 2018. Pronunciation variants and ASR of colloquial speech: A case study on Czech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Joseph Mariani, Gil Francopoulo, and Patrick Paroubek. 2019a. The nlp4nlp corpus (i): 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3.

Joseph Mariani, Gil Francopoulo, Patrick Paroubek, and Frédéric Vernier. 2019b. The nlp4nlp corpus (ii): 50 years of research in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3.

Pedro Martín-Martín. 2008. The mitigation of scientific claims in research papers: A comparative study. *IJES, International journal of english studies, ISSN 1578-7044, Vol. 8, Nº. 2, 2008 (Ejemplar dedicado a: Academic Writing: The Role of Different Rhetorical Conventions), pags. 133-152*, 8.

Lakshmi Balachandran Nair and Michael Gibbert. 2016. What makes a 'good' title and (how) does it matter for citations? a review and general model of article title attributes in management science. *Scientometrics*, 107(3):1331–1359.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Franz Josef Och and Hermann Ney. 2001. What can machine translation learn from speech recognition? In *Workshop on MT2010: Towards a Road Map for MT*, Santiago de Compostela, Spain.

Gregory S. Patience, Daria C. Boffito, and Paul A. Patience. 2015. How do you write and present research well? *The Canadian Journal of Chemical Engineering*, 93(10):1693–1696.

Seema Rawat and Sanjay Meena. 2014. Publish or perish: Where are we heading? *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, 19(2):87–89.

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, page 1–13.

Louis Rosenberg, Gregg Willcox, and Hans Schumann. 2023. Towards collective superintelligence, a pilot study. In *2023 International Conference on Human-Centered Cognitive Systems (HCCS)*, page 1–6, Cardiff, United Kingdom. IEEE.

Timo Schrader, Teresa Bürkle, Sophie Henning, Sherry Tan, Matteo Finco, Stefan Grünewald, Maira Indrikova, Felix Hildebrand, and Annemarie Friedrich. 2023. Mulms-az: An argumentative zoning dataset for the materials science domain. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, page 1–15, Toronto, Canada. Association for Computational Linguistics.

Sandeep Sharma and Jayne E. Harrison. 2006. Structured abstracts: do they improve the quality of information in abstracts? *American Journal of Orthodontics and Dentofacial Orthopedics: Official Publication of the American Association of Orthodontists, Its Constituent Societies, and the American Board of Orthodontics*, 130(4):523–530.

Luciana B. Sollaci and Mauricio G. Pereira. 2004. The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. *Journal of the Medical Library Association*, 92(3):364–371.

John Swales. 1981. *Aspects of Article Introductions*. Language Studies Unit, University of Aston in Birmingham. Google-Books-ID: Gok7NAAACAAJ.

John M Swales. 1990. *Genre analysis*. Cambridge university press.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore.

Dorothea K. Thompson. 1993. Arguing for experimental "facts" in science: A study of research article results sections in biochemistry. *Written Communication*, 10(1):106–128.

Kosuke Yamada, Tsutomu Hirao, Ryohei Sasano, Koichi Takeda, and Masaaki Nagata. 2020. Sequential span classification with neural semi-markov crfs for biomedical abstracts. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 871–877, Online. Association for Computational Linguistics.

12. Language Resource References

arXiv.org submitters. 2024. *arXiv Dataset*. Kaggle.

Bird, Steven and Dale, Robert and Dorr, Bonnie and Gibson, Bryan and Joseph, Mark and Kan, Min-Yen and Lee, Dongwon and Powley, Brett and Radev, Dragomir and Tan, Yee Fan. 2008. *The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics*. European Language Resources Association (ELRA).

Radev, Dragomir R. and Muthukrishnan, Pradeep and Qazvinian, Vahed. 2009. *The ACL Anthology Network Corpus*. Association for Computational Linguistics.

Rohatgi, Shaurya and Qin, Yanxia and Aw, Benjamin and Unnithan, Niranjana and Kan, Min-Yen. 2023. *The ACL OCL Corpus: Advancing Open Science in Computational Linguistics*. Association for Computational Linguistics.

Saier, Tarek and Krause, Johan and Färber, Michael. 2023. *unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network*.

A. Taxonomy constitution

We provide in Table 6 some details about the annotation procedure described in Section 4.1.

phase	#claim categories	annotators	#annotated sentences	#annotated articles	$\alpha(\uparrow)$	$\kappa(\uparrow)$ (min-max)
1	5	a1, a2, a3, a4	987 (a1, a2) 246 (a3, a4)	10 (a1, a2) 4 (a3, a4)	0.58	0.09-0.70
2	5	a1, a2, a5, a6	176	2	0.67	0.62-0.73
3	8	a1, a2	622	4	0.57	0.57
4	8	a1, a2	289	2	**0.81**	**0.81**

Table 6: Statistics of the different annotation phases for constituting our taxonomy of claims. Six annotators took part in the discussion to refine categories. We stabilized the categories at phase 4, where the agreement was deemed sufficient. α: Krippendorff's alpha (Krippendorff, 2011, 2013), κ: pairwise Cohen's kappa (Cohen, 1960).