# UnScientify: Detecting Scientific Uncertainty in Scholarly Full Text

Panggih Kusuma Ningrum, Philipp Mayr, Iana Atanassova

# UnScientify: Detecting Scientific Uncertainty in Scholarly Full Text

Panggih Kusuma Ningrum[1,*], Philipp Mayr[2] and Iana Atanassova[1,3]

[1]Université de Franche-Comté, CRIT, F-25000 Besancon, France
[2]GESIS — Leibniz Institute for the Social Sciences, Cologne, Germany
[3]Institut Universitaire de France (IUF), France

Abstract

This demo paper presents UnScientify (https://bit.ly/unscientify-demo), an interactive system designed to detect scientific uncertainty in scholarly full text. The system utilizes a weakly supervised technique that employs a fine-grained annotation scheme to identify verbally formulated uncertainty at the sentence level in scientific texts. The pipeline for the system includes a combination of pattern matching, complex sentence checking, and authorial reference checking. Our approach automates labeling and annotation tasks for scientific uncertainty identification, taking into account different types of scientific uncertainty, that can serve various applications such as information retrieval, text mining, and scholarly document processing. Additionally, UnScientify provides interpretable results, aiding in the comprehension of identified instances of scientific uncertainty in text.

**Keywords**

Scholarly document processing, text mining, scientific uncertainty, fine-grained annotation, pattern matching, label automation, authorial reference,

## 1. Introduction

Uncertainty is an inherent part of scientific research, as the very nature of scientific inquiry involves posing questions, developing hypotheses, and testing them using empirical evidence. Despite the best efforts of scientists to control for extraneous variables and obtain accurate measurements, there is always a certain degree of uncertainty associated with any scientific findings. This uncertainty can arise from a variety of sources, such as measurement error, sampling bias, or limitations in experimental design. Consequently, researchers resort to various strategies to manage and mitigate uncertainty when presenting their findings in academic articles. These may include using language that is overly definitive or hedging their claims with qualifiers such as "presumably" or "possible" [1].

The identification of Scientific Uncertainty (SU) in scientific text is a crucial task that can provide insights into the reliability and validity of scientific claims, help in making informed decisions, and identify areas for further investigation. Besides, detecting uncertainty has become a significant aspect of the peer-review process, which serves as a gatekeeper for the dissemination of scientific knowledge. However, the identification of scientific uncertainty in text is a complex task that requires expertise in linguistics and scientific knowledge, and is often time-consuming and labor-intensive. The primary issue stems from the fact that handling unstructured textual data in scientific literature is complicated. Previous research has mainly focused on identifying a specific set of uncertainty cues and markers in scientific articles, using a particular section of the text, such as the abstract [2] or the full text [3, 4]. These studies have helped expand the vocabulary and lexicon associated with uncertainty. However, their practical application is often inaccurate because of the intricate nature of natural language.

More sophisticated automation techniques such as machine learning and deep learning have undoubted potential for dealing with Natural Language Processing (NLP) tasks. However, the task of scientific uncertainty identification is challenging due to several factors. Firstly, there is a scarcity of available extensively annotated corpus that can be used by such techniques for scientific uncertainty identification. At present, certain corpora are limited in their scope as they only capture a particular type of uncertainty within a specific domain. For example, the BioScope corpus concentrates solely on negation or uncertainty in biological scientific abstracts [2], while the FACTBANK corpus is designed to identify the veracity or factuality of event mentions in text [5]. Similarly, the Genia Event corpus is restricted to the annotation of biological events with negation [6]. Therefore, there is a need for more diverse corpora that capture a wider range

*Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), June 26, 2023, Santa Fe, New Mexico, USA and Online*
*Corresponding author.
panggih_kusuma.ningrum@univ-fcomte.fr (P. K. Ningrum); philipp.mayr@gesis.org (P. Mayr); iana.atanassova@univ-fcomte.fr (I. Atanassova)
0000-0002-8630-6603 (P. K. Ningrum); 0000-0002-6656-1658 (P. Mayr); 0000-0003-3571-4006 (I. Atanassova)

of uncertainty types and domains, to facilitate a more comprehensive understanding of uncertainty in natural language processing.

Secondly, identifying scientific uncertainty in text involves complex linguistic features as it is often conveyed through a combination of linguistic cues, including the use of modal verbs (e.g. may, could, might), hedging devices (e.g. seems, appears, suggests), and epistemic adverbs (e.g. possibly, probably, perhaps) [7, 8]. Identifying such linguistic markers of uncertainty is not always straightforward, as they can be expressed in a variety of ways depending on the writing style or stance of the scientist.

Another challenge concerns scientists' discourse in scientific writing. A typical scientific text contains various statements and information which not only discuss the current or present study but also the former studies [9]. While writing the article, scientists can use uncertainty claims from other studies as a rhetorical tool to persuade others or to describe and organize some state of knowledge. As a result, distinguishing the reference of the uncertainty feature – whether the statement actually demonstrates uncertainty in the current study or in the former study, is a crucial factor in better understanding the context of scientific uncertainty. A study conducted by Bongelli et al. [8] is one of few that was aware of this concern. In more detail, this study only focused on the certainty and uncertainty expressed by the speakers/writers in the here-and-now of communication and excluded those that were expressed by the other party.

To overcome these challenges, we propose a weakly supervised technique that employs a fine-grained annotation scheme to construct a system for scientific uncertainty identification from scientific text focusing on the sentence level. Our approach can be used to automate labeling or annotating tasks for scientific uncertainty identification. Moreover, our annotation scheme provides interpretable results, which can aid in the understanding of the identified instances of scientific uncertainty in text. We anticipate that our approach will contribute to the development of more accurate and efficient scientific uncertainty identification systems, and facilitate the analysis and interpretation of scholarly documents in NLP.

## 2. Data

The present study employs three annotated corpora as the training set. These corpora consist of 59 journals from four different disciplines: Medicine, Biochemistry, Genetics & Molecular Biology, Multidisciplinary, and Empirical Social Science[1] which represent Science, Technology, and Medicine (STM) as well as Social Sciences and Humanities (SSH). The corpora consist of 1001 randomly selected English sentences from 312 articles across 59 journals. These sentences were annotated to identify uncertainty expressions and authorial references. By utilizing multiple corpora from different disciplines, this study aims to capture a diverse range of uncertainty expressions and improve the generalizability of the results. Table 1 illustrates the distribution of the data in the corpora and Table 2 shows the sample of annotated sentences.

## 3. Approach

Identifying scientific uncertainty in academic texts is a complex task due to various reasons. Previous research indicates that relying solely on cues or markers such as hedging words or modal verbs may not accurately identify scientific uncertainty [10]. The natural language and writing styles used by scientists, along with variations in domain-specific terminology, add to the complexity of identifying uncertainty in scientific text. Moreover, the lack of clear boundaries for expressions of uncertainty makes n-gram-based approaches too inflexible to capture the various forms and expressions of uncertainty in scientific language. To address these limitations, our research proposes a fine-grained annotation scheme for identifying uncertainty in scientific texts.

### 3.1. Fine-grained SU annotation scheme and patterns formulation

The present study adopts a span-based approach for identifying scientific uncertainty in academic text. Rather than relying solely on linguistic cues, the scheme classifies spans of text into several groups based on their linguistic features, including Part of Speech (POS) tags, morphology, and dependency. The scheme is also informed by a comprehensive analysis of scientific language, allowing for a more nuanced and accurate understanding of uncertainty expression.

During the annotation process, a list of annotated spans was created and classified into twelve groups of scientific uncertainty (SU) patterns based on their semantic meaning and characteristics. The groups include conditional expressions, hypotheses, predictions, and subjectivity, among others. In other words, the classification is based on the types of expressions used to convey uncertainty and the context in which they are used. Additionally, the scheme considers spans of text that signal disagreement statements as one of SU groups, despite ongoing debate regarding whether disagreement expressions should be considered as such. The justification for this approach is rooted in the idea that uncertainty in

[1] All social science articles are from SSOAR (https://www.ssoar.info/); we selected articles from 53 social science journals indexed in SSOAR.

| Discipline | Journal | Articles | Sentences |
|---|---|---|---|
| Medicine | BMC Med | 51 | 95 |
| | Cell Mol Gastroenterol Hepatol | 25 | 36 |
| Biochemistry, Genetics & Molecular Biology | Nucleic Acids Res | 52 | 63 |
| | Cell Rep Med | 22 | 48 |
| Multidisciplinary | Nature | 34 | 57 |
| | PLoS One | 42 | 55 |
| Empirical Social Science | SSOAR (53 journals) | 86 | 647 |

**Table 1**
Corpora description

| Sentence | SU Check | Authorial Ref. |
|---|---|---|
| It is possible that corticosteroids prevent some acute gastrointestinal complications. | Yes | Author(s) |
| However, we find no evidence to support this hypothesis either. | No | - |
| But, how this kind of coverage might influence the "we" feeling among Europeans, still remains somehow an open question. | Yes | Author(s) |
| Previous meta-analyses have shown a significant benefit for NaHCO3 in comparison to normal saline (NS) infusion [6,7], although they highlighted the possibility of publication bias. | Yes | Former/Prev. Study(s) |

**Table 2**
Samples of annotated sentences

research can stem from conflicting information or data, where multiple sources provide contradictory knowledge [11]. This type of uncertainty cannot be reduced by increasing the amount of information. Once the annotated spans are classified, Scientific Uncertainty Span Patterns (SUSP) are formulated based on the word patterns of each span and its linguistic features. Figure 1 illustrates the output from the spans annotation process.

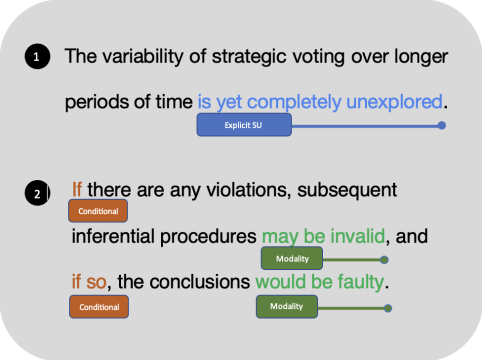



**Figure 1:** Two annotated sentences with SU expressions. Samples of output from span annotation process are shown in different colours based on their SU Pattern Group.

Figure 1 shows the application of span annotation to identify scientific uncertainty in each sentence. Each span is assigned a label corresponding to its SU pattern group. It should be noted that a sentence can have multiple labels assigned to different SU pattern groups, as seen in the second example, where labels for both conditional expression and modality are present. This feature of our annotation scheme enables the identification of complex expressions of uncertainty in scientific text. Table 3 shows more details about the list of SU pattern groups and samples from each group and more detailed information about the pattern formulation process can be seen in the demo's documentation [2].

## 3.2. Authorial Reference Checking

Authorial reference is crucial in scientific writing to provide context, especially when identifying scientific uncertainty. It helps to indicate the authorship of the argument and distinguish between the claims of the author and those of others. This can be achieved through various styles of authorial reference, such as in-text citations, reference or co-reference [12]. Additionally, there are disciplinary variations in both the frequency and use of personal and impersonal authorial references [13].

Proper attribution of uncertain claims is important to determine their origin and evaluate the credibility of the argument. For instance, when stating a hypothesis, it is essential to indicate whether it is the author's hypothesis or cited from another source. This helps the reader to assess the level of uncertainty associated with the hypothesis.

In the present study, the authorial reference of each

[2]Demo's documentation: https://bit.ly/unscientify-demo

| No | Pattern Group | Description | Examples |
|---|---|---|---|
| 1 | Explicit SU | Explicit SU group displays expressions with obvious scientific uncertainty keywords, indicating direct and explicit uncertainty expression | 1) In addition, the role of the public **is often unclear**.<br>2) ... the functional relevance of G4 in vivo in mammalian cells **remains controversial**. |
| 2 | Modality | The modality group comprises expressions that indicate uncertainty through the use of modal language | 1) Different voters **might have** different interpretations about ...<br>2) There **may also be** behavioral effects. |
| 3 | Conditional Expression | The conditional expression group includes expressions that indicate uncertainty by presenting a condition or circumstance that must be met for a certain outcome to occur | 1) **If** persons perceive the media as hostile, **it is probable that** the mere-exposure effect is weakened thus we hypothesize...<br>2) **If** there are any violations, subsequent inferential procedures may be invalid, and **if so**, the conclusions would be faulty. |
| 4 | Hypothesis | The hypothesis group encompasses expressions that indicate uncertainty by proposing a tentative explanation or assumption that requires further testing and verification to be confirmed | 1) **Hypotheses** predict that aggregate support for markets should be stronger...<br>2) **We assume** that post-materialistic individuals may have differing attitudes towards doctors than those... |
| 5 | Prediction | The prediction group comprises expressions that indicate uncertainty by proposing a forecast or projection that may or may not come to fruition, thereby introducing an element of uncertainty | 1) In July 2017, the National Grid's Future Energy Scenarios **projected that** the UK government...<br>2) Since aging leads to decreased Sir2, we **predicted that**, in young cells... |
| 6 | Interrogative Expression | The interrogative expression group includes expressions that indicate uncertainty by posing a question or series of questions, which may suggest doubt or uncertainty about a particular concept or phenomenon | 1) The study aims to determine **whether** the observed results can be replicated across different populations.<br>2) ...this research literature has also contested **whether or not** citizens' knowledge about these issues is accurate... |
| 7 | Non-generalizable statement | The non-generalizable statement group expresses uncertainty with limited scope or applicability, which may not represent a broader context or population | 1) Our study ... thus **cannot be directly generalized** to low-income nations nor extrapolated into the long-term future.<br>2) ...estimates **may not be generalisable** to women in other to women in other ancestry groups... |
| 8 | Adverbial SU | The scientific uncertainty group includes adverbs that modify or shift the sentence's meaning, introducing uncertainty | 1) ...direct and indirect readout during the transition from search to recognition mode is **poorly** understood.<br>2) It will be **quite** certain that they belong to the subpopulation of gender heterogenous... |
| 9 | Negation | The negation group comprises expressions that indicate uncertainty through the use of negation which may alter the meaning of the sentence and introduce an element of uncertainty | 1) The identity of C34 modification in... is **not clear**.<br>2) There was **no consistent** evidence for a causal relationship between age at menarche and lifetime number of sexual partners... |
| 10 | Subjectivity | The subjectivity group includes expressions indicating uncertainty through subjective language like opinions, beliefs, or personal experiences | 1) **We believe that** there are good reasons for voters to care about...<br>2) **To our knowledge**, this is the first study to provide global... |
| 11 | Conjectural | The conjectural group expresses uncertainty through conjecture or speculation, using guessing or suppositions without concrete evidence | 1) This belief **seems to be** typical for moderate religiosity.<br>2) Better performance **seems to be linked** to life satisfaction... |
| 12 | Disagreement | The disagreement group includes expressions that express uncertainty through disagreement or contradiction, often indicating opposing viewpoints or conflicting evidence | 1) **In contrast to previous studies**, our results did not show a significant effect...<br>2) **On the one hand**, some researchers argue that the use of technology in the classroom can enhance... |

**Table 3**
SU Pattern Groups and examples of annotated sentences with SU spans written in bold

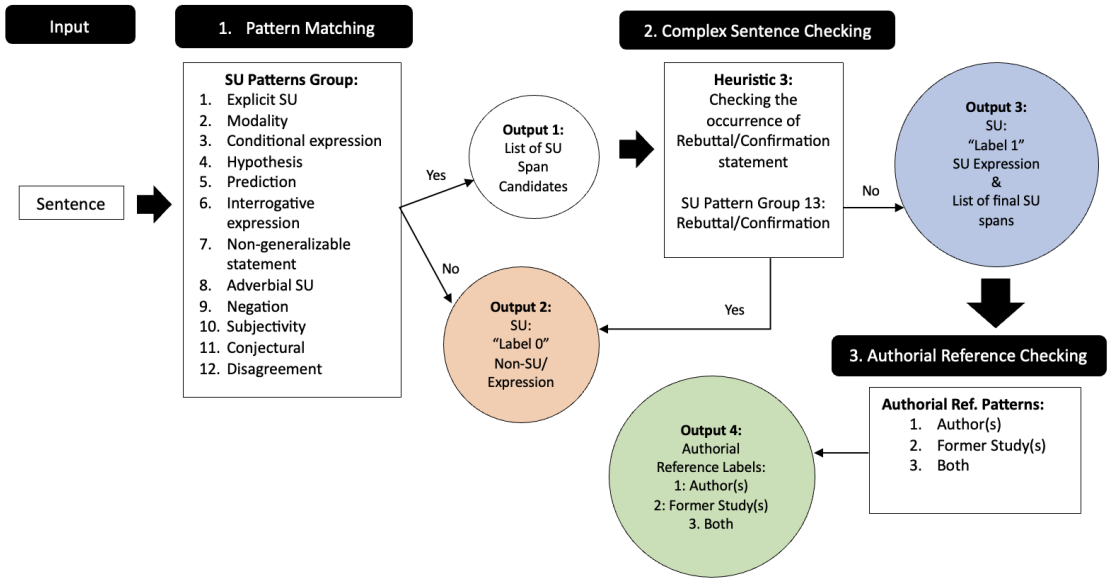



**Figure 2:** Scientific Uncertainty (SU) expression identification workflow

sentence was annotated based on the citation & co-citation patterns, and the use of personal & impersonal authorial references. Furthermore, sentences were labeled into three groups including 1) author(s) of the present article, or 2) author(s) of previous research. The last group, 3) both, is intended to accommodate complex sentences that may refer to both the author(s) and the previous study(s). Here, we present some examples of typical authorial reference mentions in context:

1. <I/We/Our study...> <text>
2. <Author/The former study...> <text>
3. (Author) (Year) <Text>
4. <Text> (Author1, Year1; Author2, Year2 . . .)
5. <Text> [Ref-No1, Ref-No2 . . . ]

## 4. Demo System

The demo system[3] for identifying SU expressions operates at the sentence level and consists of three main components: 1) Pattern Matching, 2) Complex Sentence Checking, and 3) Authorial Reference Checking, as shown in Figure 2.

The first step, Pattern Matching, employs a list of patterns derived from 12 SU pattern groups (see Table 3). The input sentence is matched against these patterns, and if a match is found, a list of SU span candidates is generated. If there is no match, the sentence is labeled as 'Non-SU expression'. To optimize the matching process, we customized a rule-based matcher from Spacy, which considers both keyword matches and patterns and linguistic features.

The second step, Complex Sentence Checking, determines whether there are any rebuttal or confirmation statements that can cancel the uncertainty expressed in the sentence. If no such statements are detected, the system labels the sentence as 'SU Expression' and provides a list of final SU spans that provide information on the reason why a particular sentence is considered a 'SU expression'.

The third step, Authorial Reference Checking, identifies the authorship of the uncertainty expression, whether it belongs to the authors, to a previous study, or both. The output of this step is the authorial reference of the sentence.

Figure 3 provides an overview of the functioning of UnScientify. The input sentence is annotated as an SU expression, matching the 'Hypothesis group' pattern. This demonstrates that UnScientify not only detects uncertainty expressions in sentences but also provides information about which sentence elements support the outcome as well as descriptive information about why the sentence is considered an SU expression. In this case, the output identifies the sentence as an SU expression due to the occurrence of the "Hypothesis group' pattern in the sentence, indicating a tentative explanation or assumption that requires further testing for confirmation. Additionally, UnScientify checks for authorial references,

[3]The demo is publicly available on https://bit.ly/unscientify-demo.

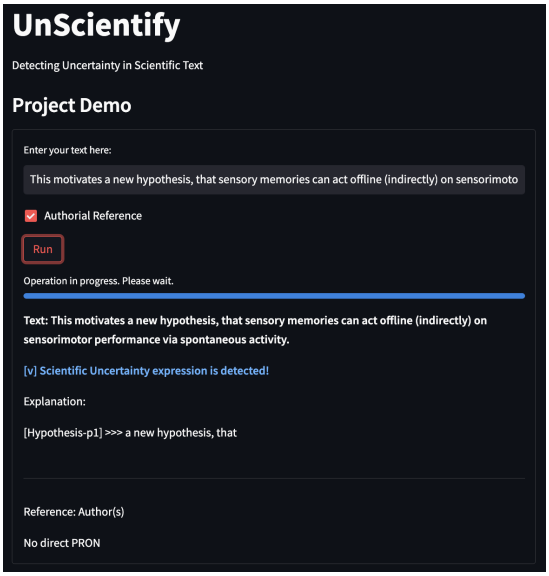



**Figure 3:** UnScientify demo interface with a sample sentence and annotation output

labeling this instance as 'Author(s)', suggesting that the sentence originates from the author rather than being cited from other sources or previous studies. As a result, it provides more contextual and interpretable results. Further demonstrations of UnScientify can be viewed in Appendix A.1.

## 5. Conclusion

Our demonstration system offers a comprehensive approach to identifying uncertainty expressions in scientific text. By utilizing pattern matching, complex sentence checking, and authorial reference checking, we provide clear and interpretable output that explains why a sentence is flagged as expressing uncertainty, addresses the element of SU expression, and verifies authorship reference.

We firmly believe that our approach holds great potential for enhancing information retrieval, text mining, and broader scientific article processing. Moreover, it lays the groundwork for further research on scientific uncertainty and epistemology. While our system currently operates at the sentence level, it can be expanded to process text at the document level.

To further enhance the UnScientify system, we acknowledge the need for improvements to identify additional dimensions of scientific uncertainty, including its nature, context, timeline, and communication characteristics. Nonetheless, we are confident that our scheme serves as a promising starting point for an in-depth exploration of how scientific knowledge is constructed and communicated.

## Acknowledgments

This research was funded by the French ANR InSciM Project (2021-2024) under grant number ANR-21-CE38-0003-01, and the Chrysalide Mobilité Internationale des Doctorants (MID) mobility grant from the University of Bourgogne Franche-Comté, France. Our appreciation extends to the GESIS – Leibniz Institute for the Social Sciences for providing the dataset and invaluable assistance, and to Nina Smirnova for her unwavering support throughout this project.

## References

[1] K. Hyland, Talking to the Academy: Forms of Hedging in Science Research Articles, Written Communication 13 (1996) 251–281. URL: https://doi.org/10.1177/0741088396013002004. doi:10.1177/0741088396013002004, publisher: SAGE Publications Inc.

[2] V. Vincze, G. Szarvas, R. Farkas, G. Móra, J. Csirik, The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes, BMC Bioinformatics 9 (2008) S9. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-S11-S9. doi:10.1186/1471-2105-9-S11-S9.

[3] B. Medlock, T. Briscoe, Weakly supervised learning for hedge classification in scientific literature, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 992–999. URL: https://aclanthology.org/P07-1125.

[4] I. Riccioni, R. Bongelli, A. Zuczkowski, Self-mention and uncertain communication in the British Medical Journal (1840-2007): The decrease of subjectivity uncertainty markers, Open Linguistics 7 (2021) 739–759. URL: https://www.degruyter.com/document/doi/10.1515/opli-2020-0179/html?lang=en. doi:10.1515/OPLI-2020-0179/MACHINEREADABLECITATION/RIS, publisher: Walter de Gruyter GmbH.

[5] R. Saurí, J. Pustejovsky, Factbank: A corpus annotated with event factuality, Language Resources and Evaluation 43 (2009) 227–268. doi:10.1007/s10579-009-9089-9.

[6] J.-D. Kim, T. Ohta, J. Tsujii, Corpus annotation for mining biomedical events from literature, BMC Bioinformatics 9 (2008) 10. URL: https:

//doi.org/10.1186/1471-2105-9-10. doi:10.1186/1471-2105-9-10.

[7] C. Chen, M. Song, G. E. Heo, A scalable and adaptive method for finding semantically equivalent cue words of uncertainty, Journal of Informetrics 12 (2018) 158–180. URL: https://linkinghub.elsevier.com/retrieve/pii/S1751157717301712. doi:10.1016/j.joi.2017.12.004.

[8] R. Bongelli, I. Riccioni, R. Burro, A. Zuczkowski, Writers' uncertainty in scientific and popular biomedical articles. A comparative analysis of the British Medical Journal and Discover Magazine, PLoS ONE 14 (2019) e0221933. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6728051/. doi:10.1371/journal.pone.0221933.

[9] S. H. Stocking, L. W. Holstein, Constructing and Reconstructing Scientific Ignorance: Ignorance Claims in Science and Journalism, Knowledge 15 (1993) 186–210. URL: https://doi.org/10.1177/107554709301500205. doi:10.1177/107554709301500205, publisher: SAGE Publications.

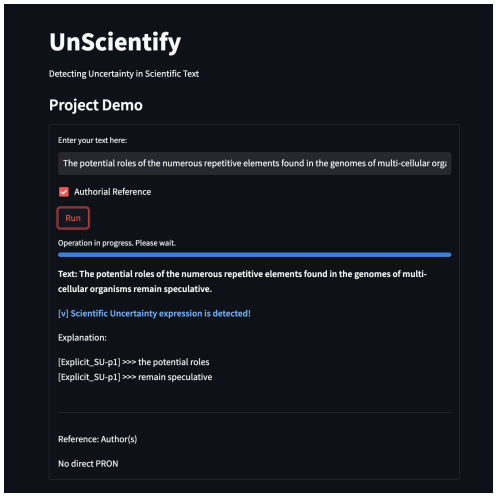
[10] P. K. Ningrum, I. Atanassova, Scientific Uncertainty: an Annotation Framework and Corpus Study in Different Disciplines, in: 19th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2023), Bloomington, Indiana, US, 2023.

[11] H. J. Zimmermann, An application-oriented view of modeling uncertainty, European Journal of Operational Research 122 (2000) 190–198. URL: https://www.sciencedirect.com/science/article/pii/S0377221799002283. doi:10.1016/S0377-2217(99)00228-3.

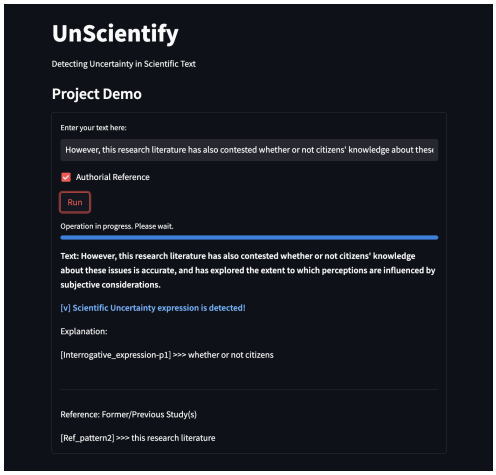
[12] B. Powley, R. Dale, Evidence-Based Information Extraction for High Accuracy Citation and Author Name Identification, 2007.

[13] M. Khedri, K. Kritsis, How do we make ourselves heard in the writing of a research article? A study of authorial references in four disciplines, Australian Journal of Linguistics 40 (2020) 194–217. URL: https://www.tandfonline.com/doi/full/10.1080/07268602.2020.1753011. doi:10.1080/07268602.2020.1753011.

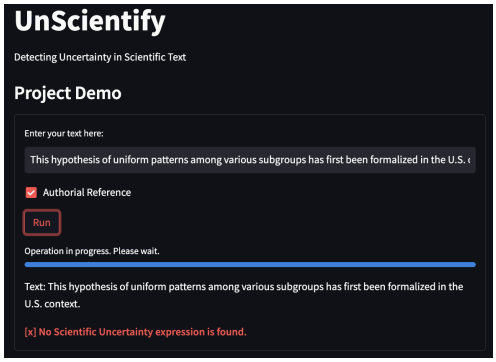
# A. Appendix



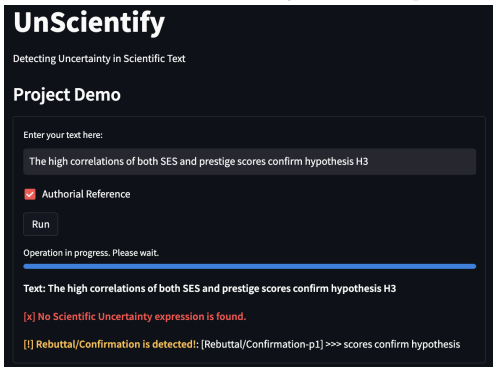

(a) Demo 1: Detecting Explicit SU with multiple SU spans




(b) Demo 2: Detecting a sentence containing an interrogative expression




(c) Demo 3: A sentence showing a Non-SU expression




(d) Demo 4: Rebuttal/Confirmation Statement Detection

**Figure A.1:** UnScientify Demos