UnScientify: Detecting Scientific Uncertainty in Scholarly Full Text

EEKE-AII 2023 Workshop, 26-27 June 2023
JCDL 2023

By
Panggih Kusuma NINGRUM (Université de Franche-Comté, CRIT, France)
Philipp MAYR (GESIS — Leibniz Institute for the Social Sciences, Germany)
Iana ATANASSOVA (Université de Franche-Comté, CRIT, France)

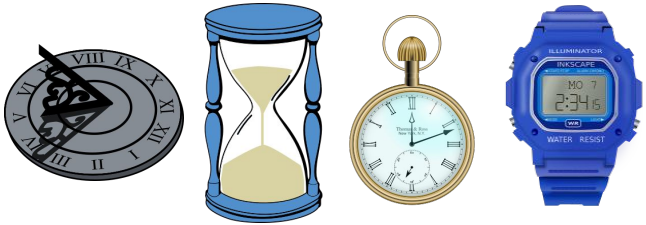
Contents

Designed by Freepik

1. Background

Don’t scientists know everything?

We can never be completely certain about the future, either in everyday life, or in science.



Source: https://openclipart.org/

Scientists face uncertainty at numerous stages of their research process (Cordner and Brown, 2013)

1. Background

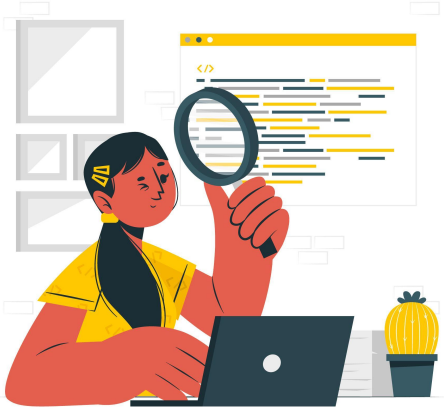
Consequently...

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" (Hyland, 1996)

Designed by Freepik

2. Problem Statement

Why is detecting Scientific uncertainty a big deal?



Designed by Freepik

- provide insights into the reliability and validity of scientific claims, help in making informed decisions, and identify areas for further investigation
- become a significant aspect of the peer-review process, which serves as a gatekeeper for the dissemination of scientific knowledge

2. Problem Statement

- SU identification requires expertise in linguistics and scientific knowledge, time-consuming and labor-intensive.
- A scarcity of available extensively annotated corpus - certain corpora are limited in their scope as they only capture a particular type of uncertainty within a specific domain.
- Typical scientific text contains various statements and information which not only discuss the current or present study but also the former studies (Stocking and Holstein, 1993)

Designed by Freepik

Therefore...

(+) automates labelling and annotation tasks for scientific uncertainty identification

(+) taking into account different types of scientific uncertainty

(+) provides interpretable results: also provides information about which sentence elements support the outcome & authorial reference

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.

3. Data

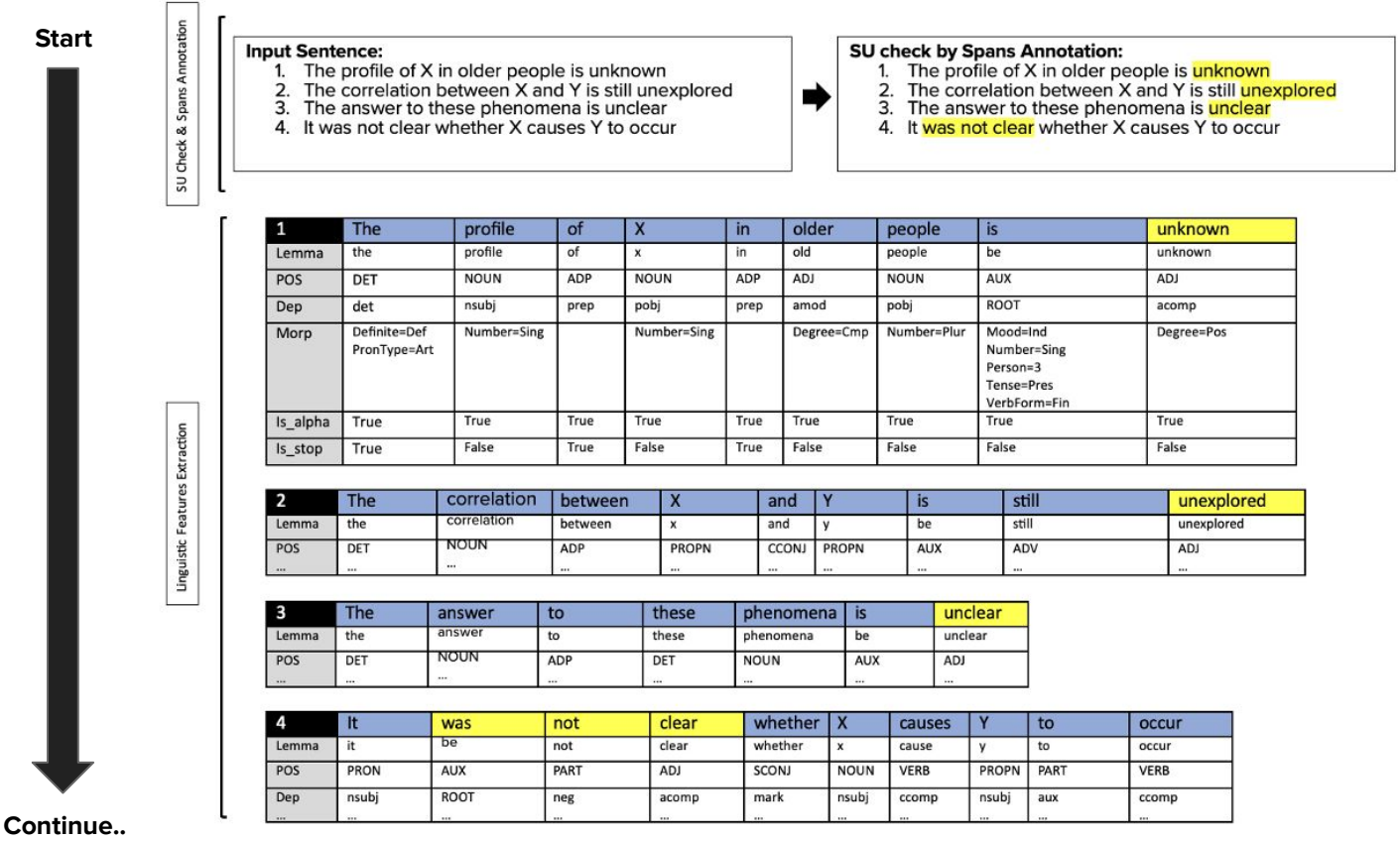
Table 1. Corpora Description (Annotated Datasets)

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 2. Samples of annotated sentences

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)

4. Approach



1	The	profile	of	X	in	older	people	is	unknown
Lemma	the	profile	of	x	in	old	people	be	unknown
POS	DET	NOUN	ADP	NOUN	ADP	ADJ	NOUN	AUX	ADJ
Dep	det	nsubj	prep	pobj	prep	amod	pobj	ROOT	acomp
Morp	Definite=Def PronType=Art	Number=Sing		Number=Sing		Degree=Cmp	Number=Plur	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	Degree=Pos
Is_alpha	True	True	True	True	True	True	True	True	True
Is_stop	True	False	True	False	True	False	False	False	False

2	The	correlation	between	X	and	Y	is	still	unexplored
Lemma	the	correlation	between	x	and	y	be	still	unexplored
POS	DET	NOUN	ADP	PROPN	CCONJ	PROPN	AUX	ADV	ADJ
...	...	...	...	...	...	...	...	...	...

3	The	answer	to	these	phenomena	is	unclear
Lemma	the	answer	to	these	phenomena	be	unclear
POS	DET	NOUN	ADP	DET	NOUN	AUX	ADJ
...	...	...	...	...	...	...	...

4	It	was	not	clear	whether	X	causes	Y	to	occur
Lemma	it	be	not	clear	whether	x	cause	y	to	occur
POS	PRON	AUX	PART	ADJ	SCONJ	NOUN	VERB	PROPN	PART	VERB
Dep	nsubj	ROOT	neg	acomp	mark	nsubj	ccomp	nsubj	aux	ccomp
...	...	...	...	...	...	...	...	...	...	...

Figure 1. SU Pattern Formulation

Continue..

Patterns Formulation

Finish

Keywords & Pattern Detection:

1. The profile of X in older people **is unknown** → <Lemma:be, dep:ROOT> <POS:ADJ>
2. The correlation between X and Y **is still unexplored** → <Lemma:be, dep:ROOT> <POS:ADV> <POS:ADJ>
3. The answer to these phenomena **is unclear** → <Lemma:be, dep:ROOT> <POS:ADJ>
4. It **was not clear** whether X causes Y to occur → <Lemma:be, dep:ROOT> <dep:neg> <POS:ADJ>

Keywords Classification

Uncertainty Keywords Dictionary

ADJ SU	ADJ SU Antonym	NOUN SU	...
unknown	known	uncertainty	...
unexplored	explored	controversy	...
unclear	clear	ambiguity	...
unsure	sure	probability	...
speculative	proven	hypothesis	...
...	...	...	...

Generating Patterns

Pattern1 : <Lemma:Be, dep:ROOT> < ! Negation > <ADV * > <POS:ADJ in ADJ SU>

Matched Sentence:

1. The profile of X in older people is unknown
2. The correlation between X and Y is still unexplored
3. The answer to these phenomena is unclear

...

m. It is unknown whether these missing data have influenced the results.

Pattern2 : <Lemma:Be, dep:ROOT> < Negation > <ADV * > <POS:ADJ in ADJ SU Antonym>

Matched Sentence:

4. It was not clear whether X causes Y to occur

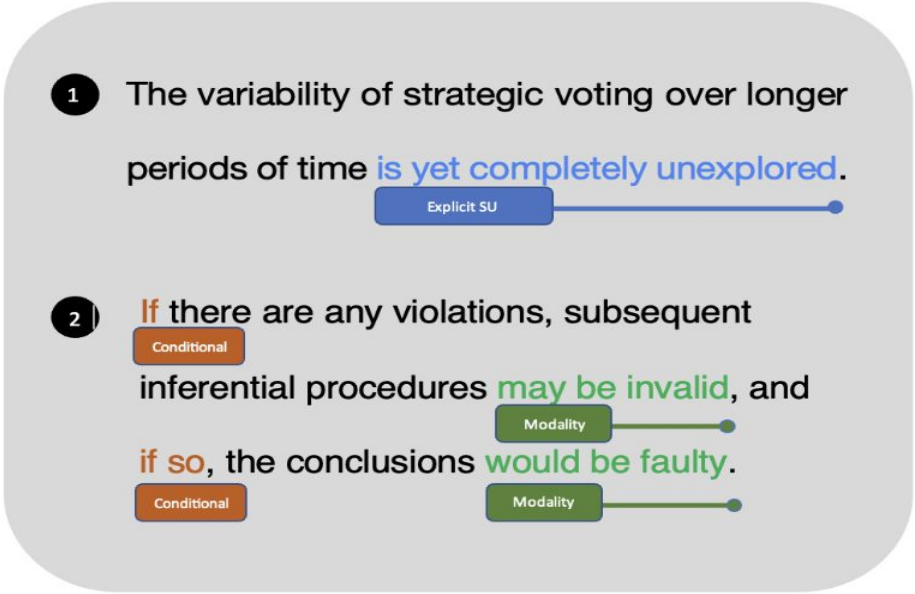
...

n. The factors contributing to the event are still not known

4. Approach

SU Patterns Group:

1. Explicit SU
2. Modality
3. Conditional expression
4. Hypothesis
5. Prediction
6. Interrogative expression
7. Non-generalizable statement
8. Adverbial SU
9. Negation
10. Subjectivity
11. Conjectural
12. Disagreement



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

5. System

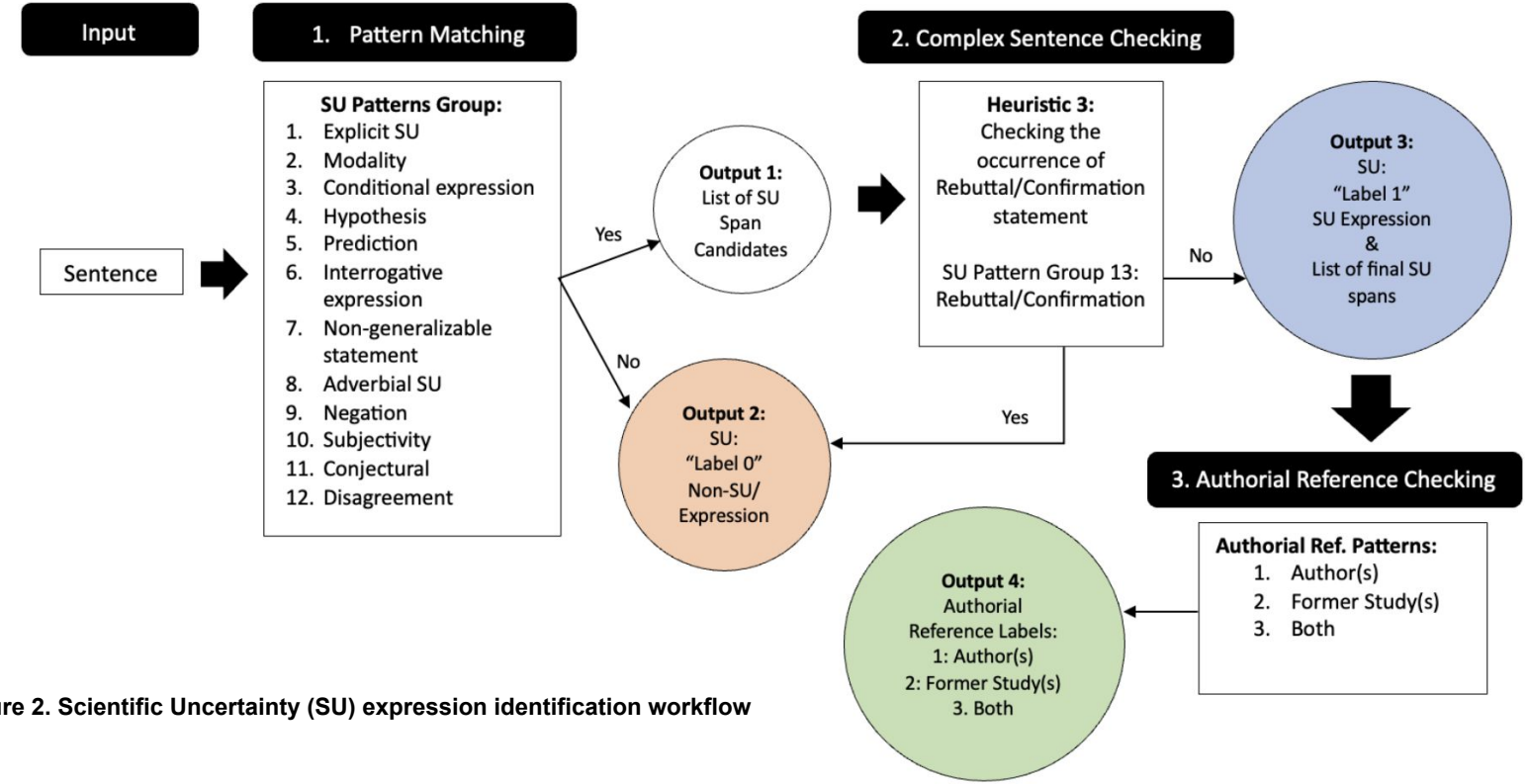


Figure 2. Scientific Uncertainty (SU) expression identification workflow

5. System

1 2

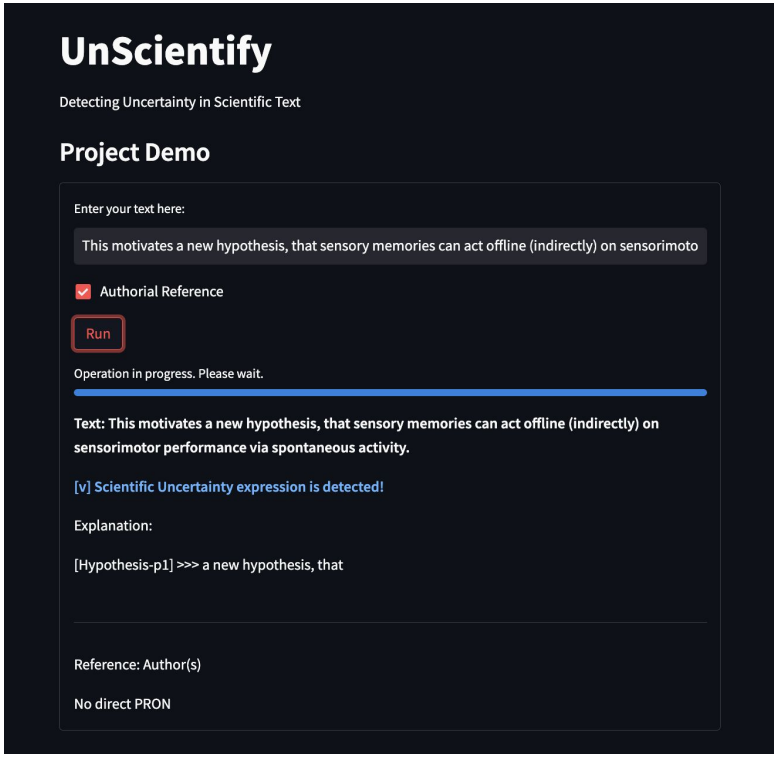
The authorial reference of each sentence was annotated based on the citation & co-citation patterns, and the use of personal & impersonal authorial references. Furthermore, sentences were labeled into three groups including:

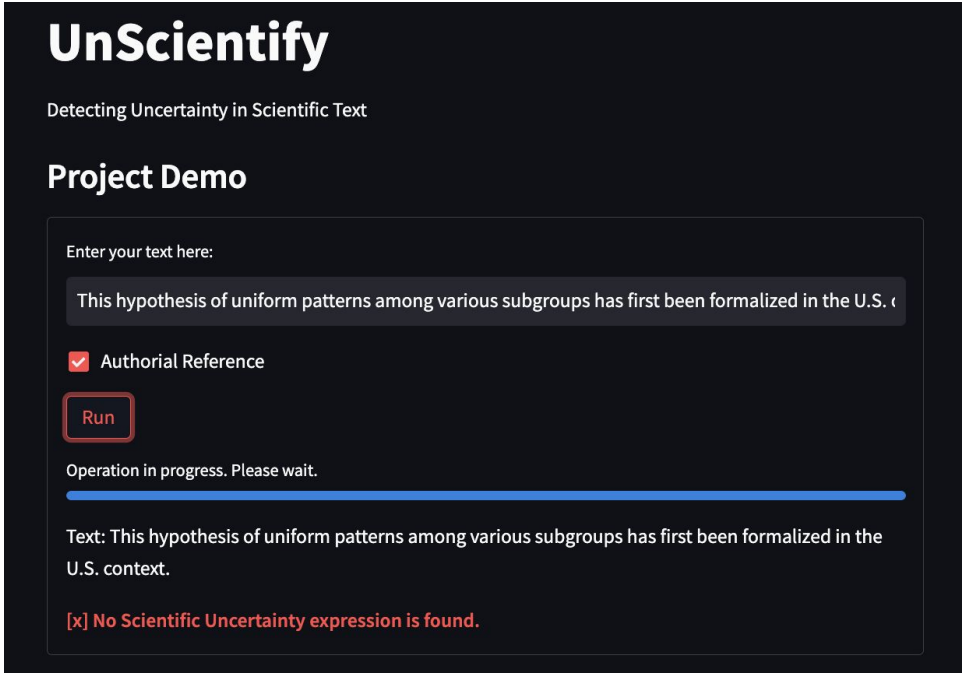
1. **Author(s) of the present article, or**
2. **Author(s) of previous research**
3. **Both, is intended to accommodate complex sentences that may refer to both the author(s) and the previous study(s).**

Samples of authorial patterns:

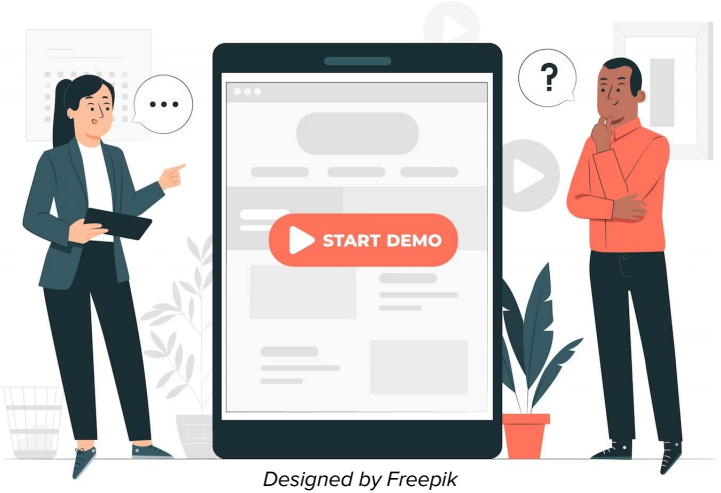
1. **<I/We/Our study…>** <text>
2. **<Author/The former study…>** <text>
3. **(Author) (Year)** <Text>
4. <Text> **(Author1, Year1; Author2, Year2 . . .)**
5. <Text> [Ref-No1, Ref-No2 . . .]

6. UnScientify Demo





6. UnScientify Demo



Designed by Freepik

https://bit.ly/unscientify-demo

7. Further Improvements

- Improvements to identify additional dimensions of scientific uncertainty, including its nature, context, timeline, and communication characteristics
- Currently UnScientify operates at the sentence level, it can be expanded to process text at the document level.

Designed by Freepik

8. References

[1] Ramona Bongelli, Ilaria Riccioni, Roberto Burro, and Andrzej Zuczkowski. 2019. Writers' uncertainty in scientific and popular biomedical articles. A comparative analysis of the British Medical Jour- nal and Discover Magazine. PLoS ONE 14, 9 (Sept. 2019), e0221933. https://doi.org/10.1371/journal.pone.0221933

[2] Chaomei Chen, Min Song, and Go Eun Heo. 2018. A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. Journal of Informetrics 12, 1 (Feb. 2018), 158–180. https: //doi.org/10.1016/j.joi.2017.12.004

[3] Ken Hyland. 1996. Talking to the Academy: Forms of Hedging in Science Research Articles. Written Communication 13, 2 (April 1996), 251–281. https://doi.org/10.1177/0741088396013002004 Publisher: SAGE Publications Inc.

[4] Mohsen Khedri and Konstantinos Kritsis. 2020. 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, 2 (April 2020), 194–217. https://doi.org/10.1080/07268602.2020.1753011

[5] Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annota- tion for mining biomedical events from literature. BMC Bioinformatics 9, 1 (Jan. 2008), 10. https://doi.org/10.1186/1471-2105-9-10

[6] Ben Medlock and Ted Briscoe. [n. d.]. Weakly Supervised Learning for Hedge Classification in Scientific Literature. ([n. d.]), 8.

[7] Panggih Kusuma Ningrum and Iana Atanassova. 2023. Scientific Un- certainty: 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.

[8] Brett Powley and Robert Dale. 2007. Evidence-Based Information Ex- traction for High Accuracy Citation and Author Name Identification.

[9] Ilaria Riccioni, Ramona Bongelli, and Andrzej Zuczkowski. 2021. Self- mention and uncertain communication in the British Medical Journal (1840-2007): The decrease of subjectivity uncertainty markers. Open Linguistics 7, 1 (Jan. 2021), 739–759. https://doi.org/10.1515/OPLI2020-0179/MACHINEREADABLECITATION/RIS Publisher: Walter de Gruyter GmbH.

[10] RoserSauríandJamesPustejovsky.2009.Factbank:Acorpusannotated with event factuality. Language Resources and Evaluation 43, 3 (Sept. 2009), 227–268. https://doi.org/10.1007/s10579-009-9089-9

[11] S. Holly Stocking and Lisa W. Holstein. 1993. Constructing and Re- constructing Scientific Ignorance: Ignorance Claims in Science and Journalism. Knowledge 15, 2 (Dec. 1993), 186–210. https://doi.org/10. 1177/107554709301500205 Publisher: SAGE Publications.

[12] Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 9, S11 (Dec. 2008), S9. https://doi.org/10.1186/1471-2105-9-S11-S9

[13] H. J. Zimmermann. 2000. An application-oriented view of modeling uncertainty. European Journal of Operational Research 122, 2 (April 2000), 190–198. https://doi.org/10.1016/S0377-2217(99)00228-3

8. Acknowledgment

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.

Thank you
Merci
Danke
Terimakasih

panggih_kusuma.ningrum@univ-fcomte.fr
Philipp.Mayr@gesis.org
iana.atanassova@univ-fcomte.fr