InSciM Progress Report 2023

By
Panggih Kusuma NINGRUM
PI: Iana ATANASSOVA

Besançon, 20 June 2023

Contents

1. Overview
2. Data
3. Current Stage
4. Results (on going)
5. Challenges & Further Improvement
6. UnScientify Demo

1. Overview

Milestone:

1. **Publications:**
   a. Ningrum, Panggih Kusuma & Atanassova, Iana. (2023). ”**Dataset for Multidisciplinary Uncertainty Mining - ver1 (Version 1)**” [Data set]. Zenodo. https://doi.org/10.5281/zenodo.8024787
   b. Ningrum, P. K., Atanassova, I. (2023) "**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.
   c. Ningrum, P. K., Mayr, P., Atanassova, I. (2023) “**UnScientify: Detecting Scientific Uncertainty in Scholarly Full Text**” In Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI + Informetrics (AII2023), part of the ACM/IEEE Joint Conference on Digital Libraries 2023, Santa Fe, New Mexico, USA, June 26 - 30, 2023

2. **Research Visiting program and collaboration with GESIS**
   Duration: 3 months (February - May 2023)
   Output: Annotated dataset in the field of empirical social science, UnScientiFy app (demo), 1 paper

2. Data

Table 1. Corpora Description & Total Number of Targeted Scientific Articles*

Discipline	Subject Area	Total Documents*
Medicine	Medicine	50.247
Non-Medicine	Arts & Humanities	77.632
	Biochemistry, Genetics & Molecular Biology	4.548
	Computer Science	11.476
	Environmental Science	25.509
	Physics & Astronomy	2.565
	Psycology	34.184
	Social Sciences	1.932
Multidisciplinary	Plos One	269.033
	Nature	4.160
	arXiv	1.700.000
Total		2.181.286

*Total data in each journal including Article. Editorial, Correction, Commentary, Corrigendum, Erratum, etc.

\+ Empirical Social Science Articles (SSOAR) - GESIS

3. Research Pipeline & Methodology D 1 2

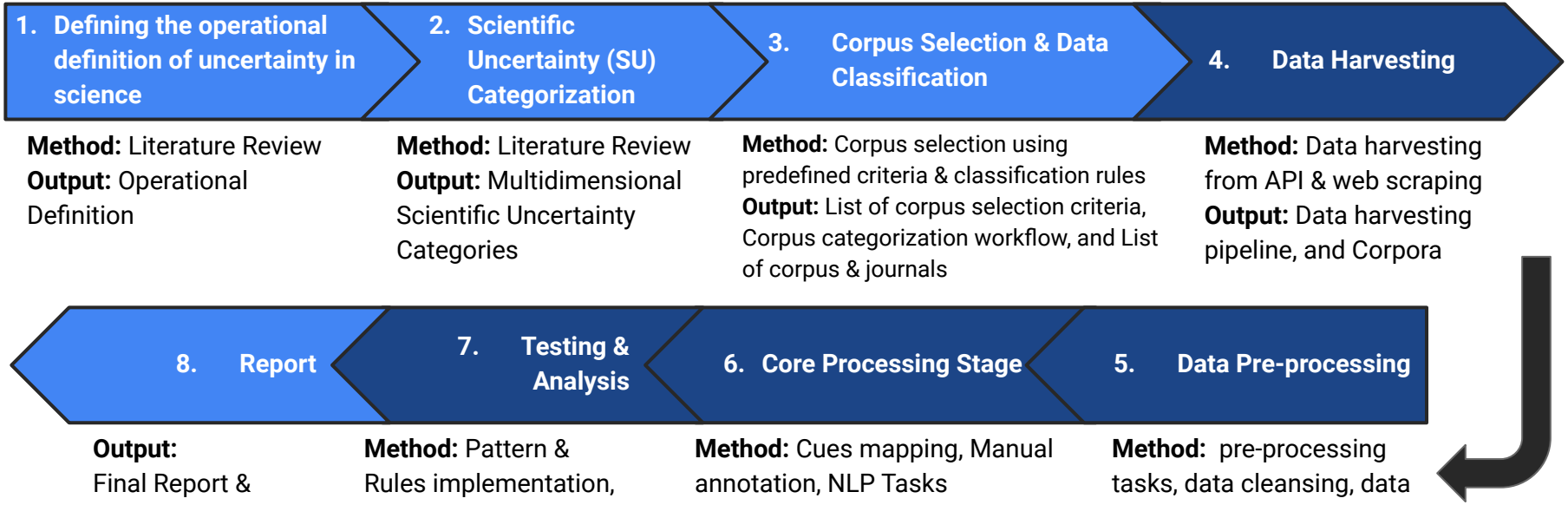


Figure 1. Core Processing Stage

3. Research Pipeline & Methodology D1 D2

Table 2. Core Processing Stage

Step	Input	Process	Methods	Output
SU Search	1. Samples 2. Uncertainty Cue List (Hyland, 1996; Chen et al. 2018; Bongelli et al. (2019) 3. 5D SU Categorization	1. Cue Mapping Process 2. Manual Search 3. Annotation	- NLP Tasks - Manual Annotation	**Output 1** SU Expressions List + Label (Y/N, Categories)
SU Keywords & Span Extraction	Output 1	1. Keywords & Span Extraction	- NLP Tasks - Manual Annotation	**Output 2** Keywords & Spans List
SU Patterns & Rules Formulation	Output 2	1. Linguistic Features Extractions 2. Syntagmatic Relation Analysis 3. Clustering	- NLP Tasks - Manual Classification	**Output 3** Patterns list **Output 4** Rules & Heuristics
Patterns & Rules Test	1. Output 3 2. Output 4 3. Annotated Testing Data	Patterns & Rules Implementation	- NLP Tasks	**Output 5** Result of Patterns & Rules Performance
Evaluation & Improvement	Output 5	Analysis & Evaluation	- Data Analysis	**Output 6** Final Results

4. Results

1. Multidimensional Scientific Uncertainty Categorization

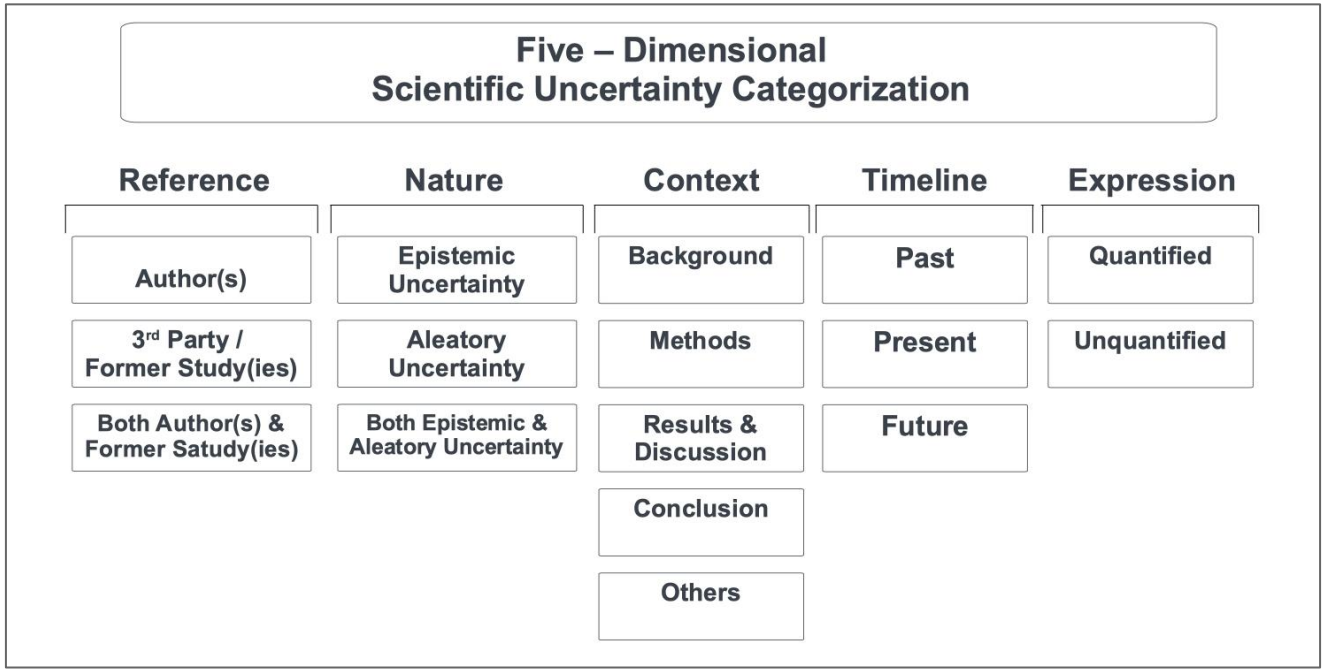


Figure 2. Scientific Uncertainty Categorization

4. Results

1. Multidimensional Scientific Uncertainty Categorization (Cont.)

Category	Description
1. Reference	**Addressing the "Who" or authorial reference** of the scientific uncertainty expression, whether it refers to the author(s) of the observed journal article or the third party or author(s) from former research. *(Stocking and Holstein, 1993)*
2. Nature	**Epistemic Uncertainty** refers to deficiencies caused by a **lack of knowledge** or information complexity. In theory, knowledge creation and learning can help to reduce this type of uncertainty. ➔ *"I am **70% sure** that..."* **Aleatory uncertainty** refers to the uncertainty resulting from inherent variability or uncertainty brought on by **probabilistic variations in a random event**. ➔ *"I **think** there is a 75% **chance** that..."*
3. Context	The context of uncertainty is the manner in which uncertainty emerges itself within the journal article. (*Friedman et al., 1999)*
4. Timeline	The relevance of time (past, present, and future) to the moment when the article was written. *(Rubin et. al. 2006)*
5. Expression	How uncertainty is delivered and communicated in text. *(Van der Bles et. al., 2018)* **Quantifiable** ➔ absolute quantitative terms, including a probability distribution or confidence interval, etc **Unquantifiable** ➔a series of caveats about the underlying sources of evidence, which can be combined into a qualitative scale

4. Results

2. Annotated Datasets

Table 4. Annotated Datasets Description

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	86	647

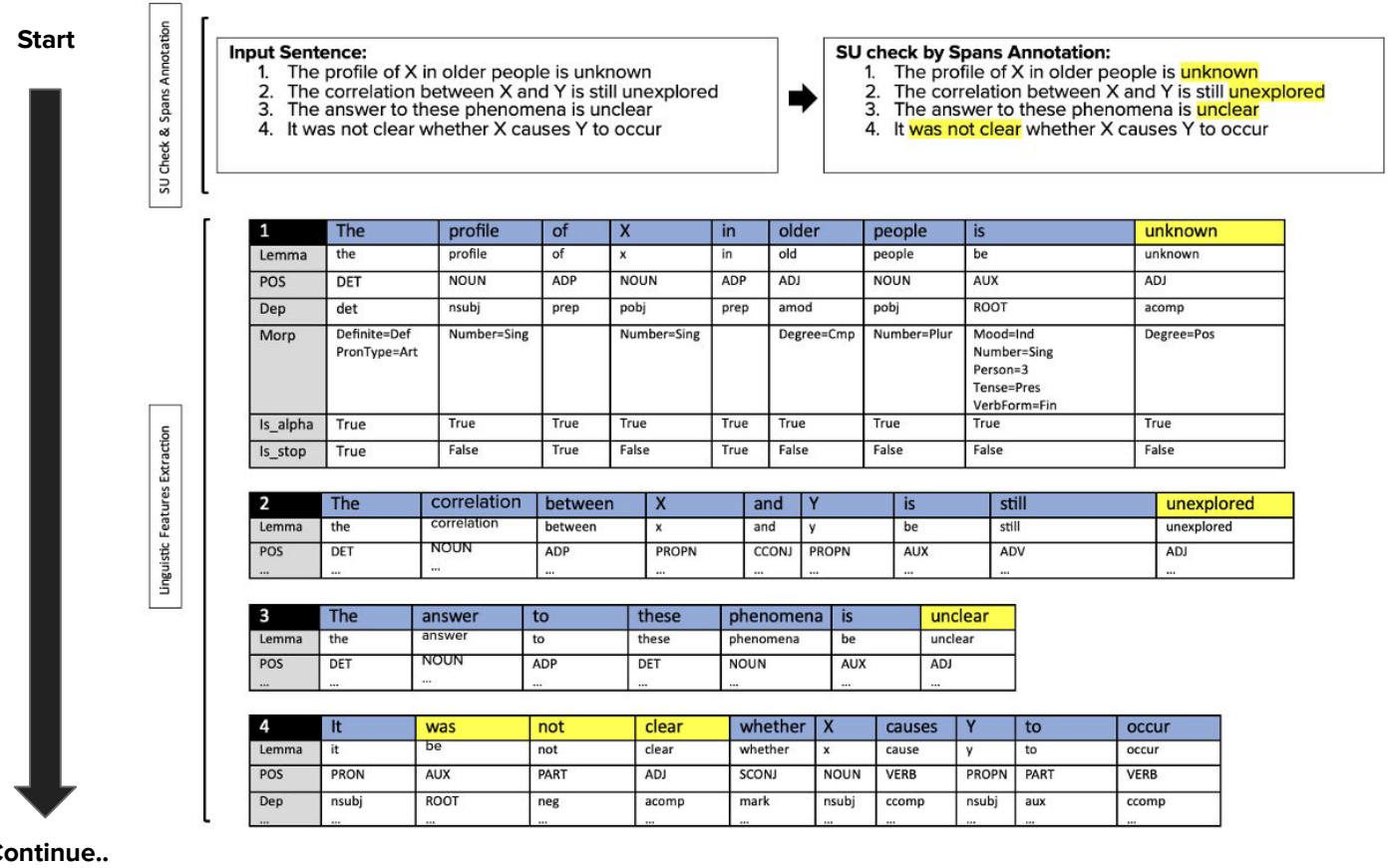
4. Results

2. Annotated Datasets

Examples of sentences and annotations

Sentence	Journal	Reference	Nature	Context	Timeline	Expression
Recent studies suggest that the African ZIKV lineage virus has higher transmissibility and pathogenicity compared to the Asian lineage strain, and infection in pregnant women may be more likely to cause total fetal loss than congenital deformities associated with the Asian lineage [15].	BMC Med	Former/Previous Study(s)	Epistemic	Background	Past	Unquantified
It is possible that corticosteroids prevent some acute gastrointestinal complications.	BMC Med	Author(s)	Aleatory	Conclusion	Present	Unquantified
Additional studies are required to further characterize pathways linking bacterial metabolites with environment-modulated mechanisms driving carcinogenesis in the colon mucosa.	Cell Mol Gastroenterol Hepatol	Author(s)	Epistemic	Results & Discussion	Future	Unquantified

3. SU Pattern Formulation



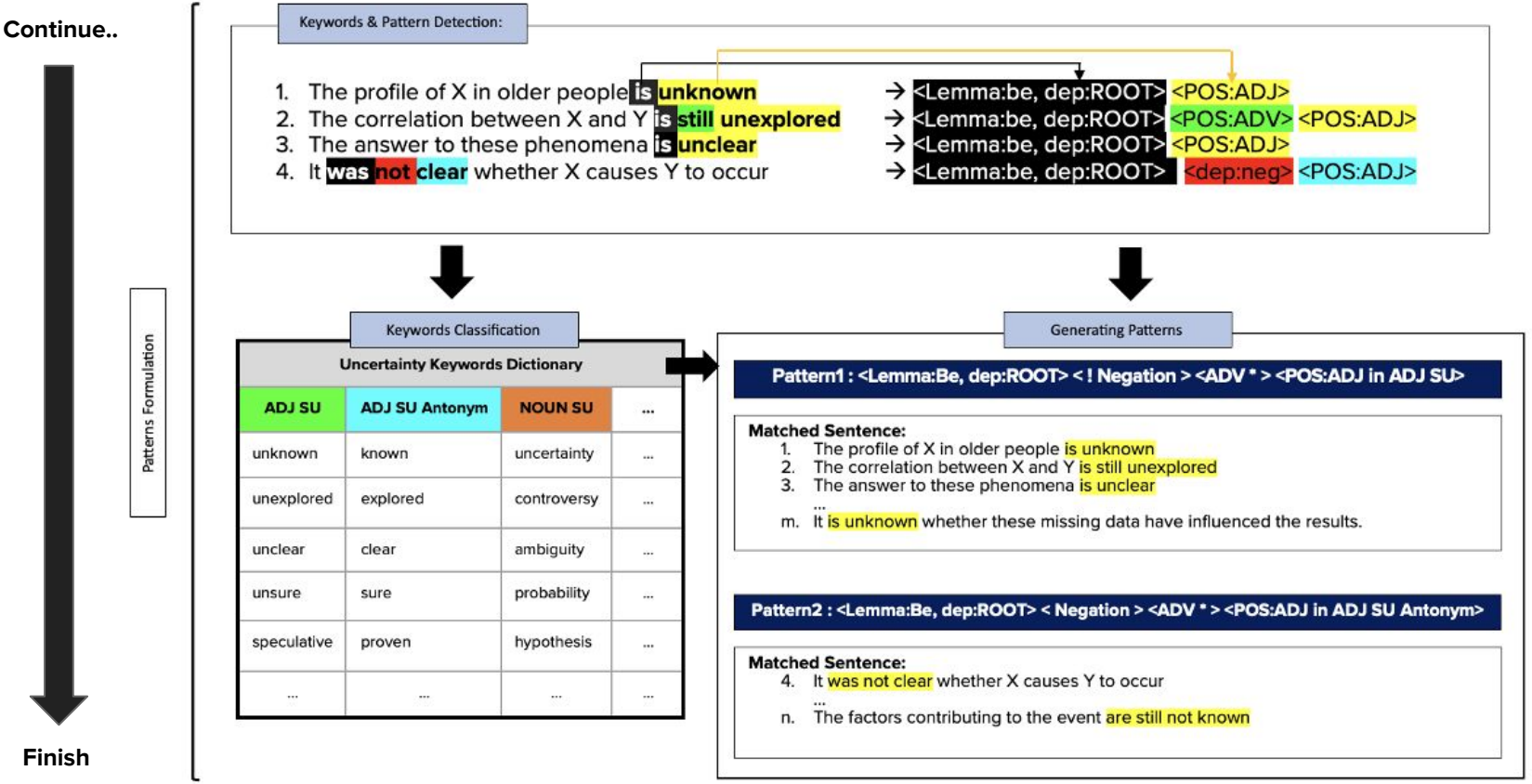
1	The	profile	of	X	in	older	people	is	unknown
Lemma	the	profile	of	x	in	old	people	be	unknown
POS	DET	NOUN	ADP	NOUN	ADP	ADJ	NOUN	AUX	ADJ
Dep	det	nsubj	prep	pobj	prep	amod	pobj	ROOT	acomp
Morp	Definite=Def PronType=Art	Number=Sing		Number=Sing		Degree=Cmp	Number=Plur	Mood=Ind Number=Sing Person=3 Tense=Pres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

3. SU Pattern Formulation (Cont.)

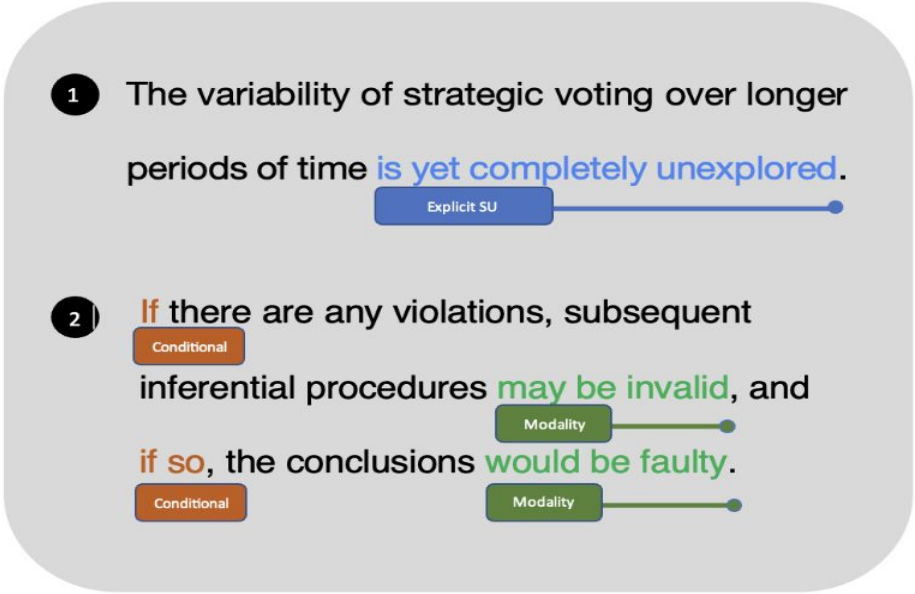


4. Results

2. SU Pattern Groups

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.

4. Results

3. Reference (Authorial Patterns)

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 . . .]

5. Challenges & Further Improvement

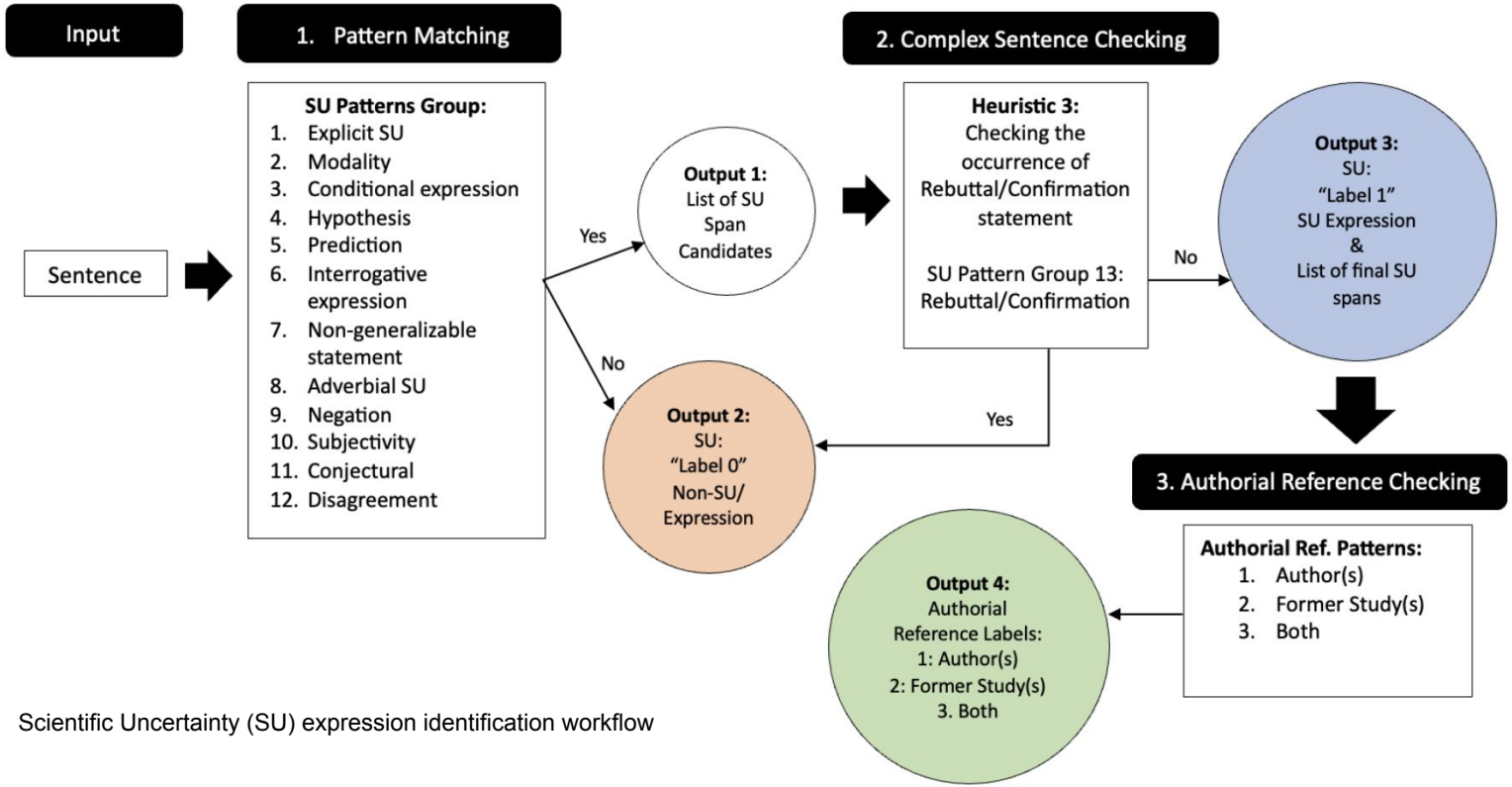
1. Challenges

- Data harvesting & processing (A new postdoc will join in about January 2024)
- Need more annotators (Interns will join in January 2024)
- Training annotators will take some times as the annotators need to have a strong fundamental knowledge about Uncertainty in Science

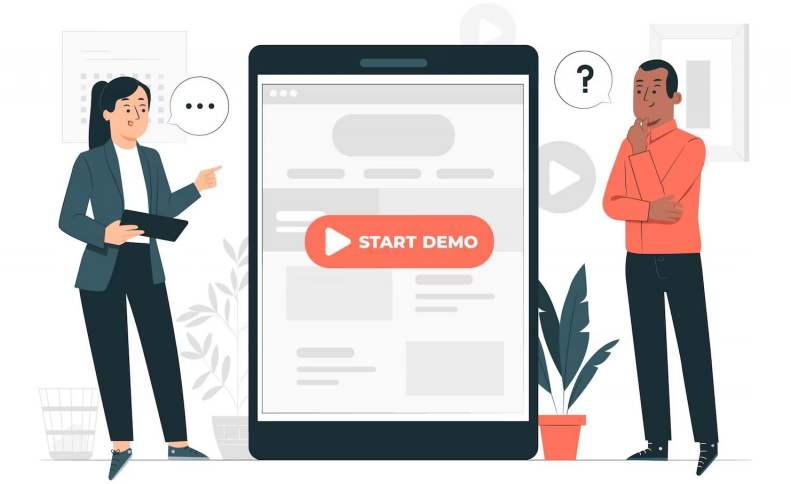
2. Further Improvement

- Continuing data harvesting & pre-processing
- Continuing Manual annotation
- Extending the process to other dimensions: Nature, Context, Timeline & Expression
- App testing & evaluation (Post Doc)

6. UnScientify Demo



Scientific Uncertainty (SU) expression identification workflow



demo