



# Categorising Scientific Uncertainty in Papers

Iana Atanassova, François-C. Rey

## ► To cite this version:

Iana Atanassova, François-C. Rey. Categorising Scientific Uncertainty in Papers. SciNLP 2021, 8 October 2021, 2nd Workshop on Natural Language Processing for Scientific Text, Oct 2021, Irvine, United States. , <https://scinlp.org>, 2021. hal-03476393

**HAL Id: hal-03476393**

**<https://hal.science/hal-03476393>**

Submitted on 16 Dec 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Categorising Scientific Uncertainty in Papers

Iana Atanassova<sup>1,2</sup> and Francois-C. Rey<sup>1</sup>

<sup>1</sup> CRIT, Université de Bourgogne Franche-Comté, France

<sup>2</sup> Institut Universitaire de France (IUF)

iana.atanassova@univ-fcomte.fr,  
francois\_claude.rey@edu.univ-fcomte.fr



## Why study uncertainty in science?

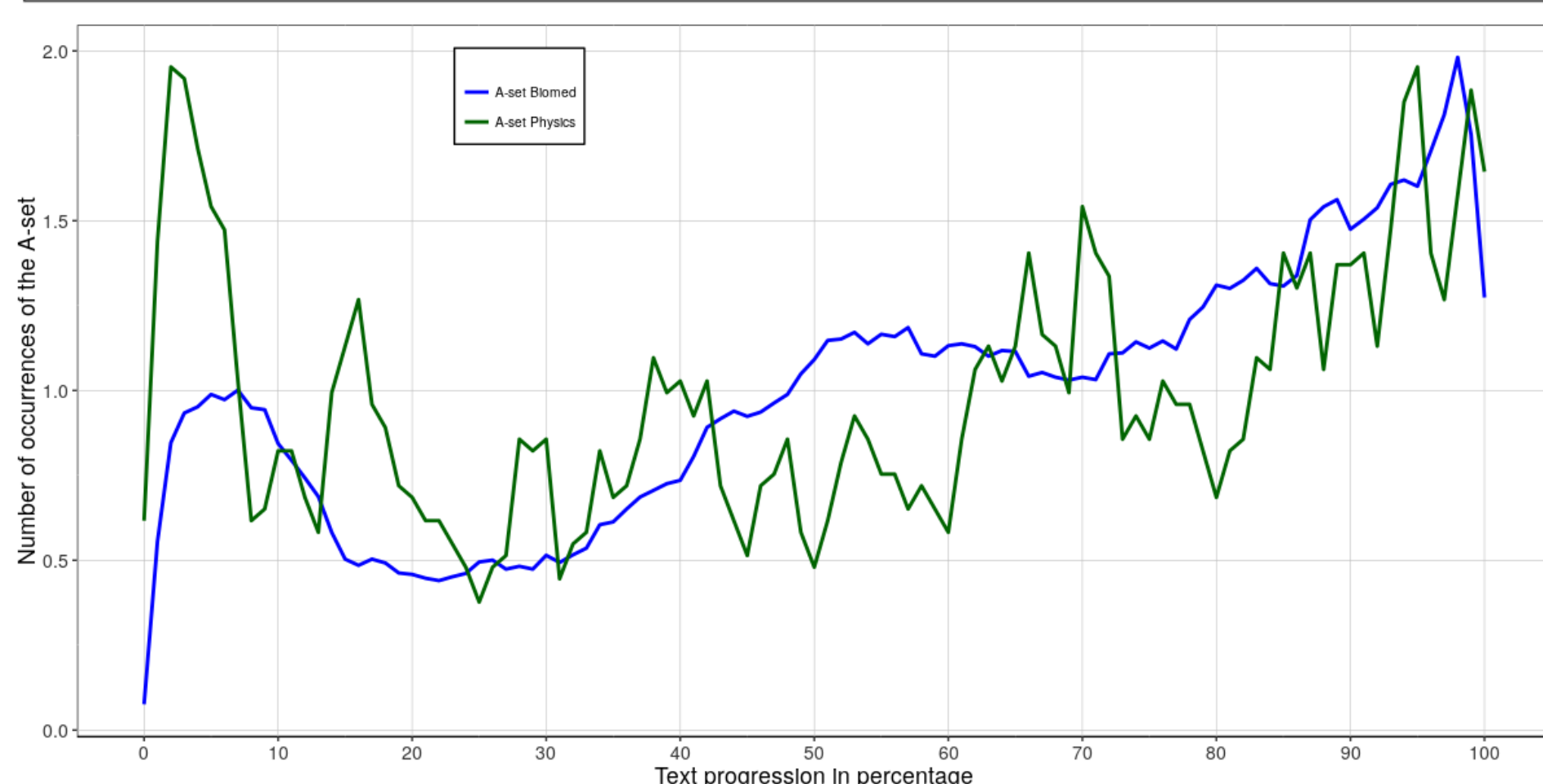
- Uncertainty in science is an integral part of the research process and an important element of discovery, and as such it is expressed in publications.
- The use of tools or observations produces a margin of error, and the use of abductive and inductive reasoning in science implies the presence of uncertainty.
- Uncertainty is specific to each discipline, linked to the object of the study, to the methodologies, or to the results.
- The expression of uncertainty in scientific articles makes use of complex linguistic properties.

Our objective is to study the expression of uncertainty in papers, in order to automatically identify and classify sentences that contain uncertainty.

## Distribution of indicators of uncertainty [1]

We have processed two datasets of papers part of PubMed OA dataset: a Biomed dataset of 9 463 papers from 7 journals, and a Physics dataset of 488 papers from 2 journals.

A-set: strong indicators of uncertainty  
raises (some) doubts about  
there is no (clear) evidence of/about  
more/further (...) studies/research/experiments /evaluation (are/is) needed to  
may enforce the concept/theory/model of/about  
it is plausible/possible/probable that  
it is difficult/impossible to draw a (general) conclusion  
we cannot be certain/sure that/if/whether  
do/does not allow determining/identifying/measuring/evaluating ... with (absolute/greater) certainty  
cannot be determined/identified/measured/evaluated ... with (absolute/greater) certainty  
we cannot state/formulate/assess with (absolute/greater) certainty



## References

- [1] Iana Atanassova, François-C. Rey, and Marc Bertin. Studying Uncertainty in Science: a distributional analysis through the IMRaD structure. *Conférence 7<sup>th</sup> International Workshop On Mining Scientific Publications (WOSP) - LREC, 7-12 May 2018, Miyazaki, Japan*, 2018.
- [2] François-C. Rey, Marc Bertin, and Iana Atanassova. Une étude de l'incertitude dans les textes scientifiques : vers la construction d'une ontologie. *Terminologie & Ontologie : Théories et Applications (TOTH), Chambéry, France*, 2018.

## Acknowledgements

Part of this research has been funded by the ANR-JCJC Project InSciM - "Modelling Uncertainty in Science".

## Examples of sentences

### Measurable uncertainty

*"In the former Soviet Union the potential for increased carbon sequestration in agricultural soils is much greater, perhaps by an order of magnitude."*

*"This population of 250 individuals has a 50% chance of extinction over the next 100 years."*

### Qualitative uncertainty

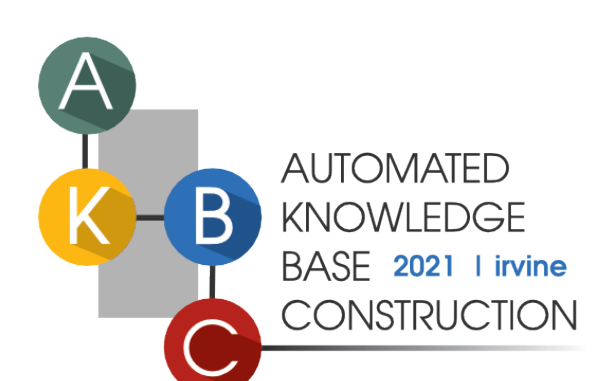
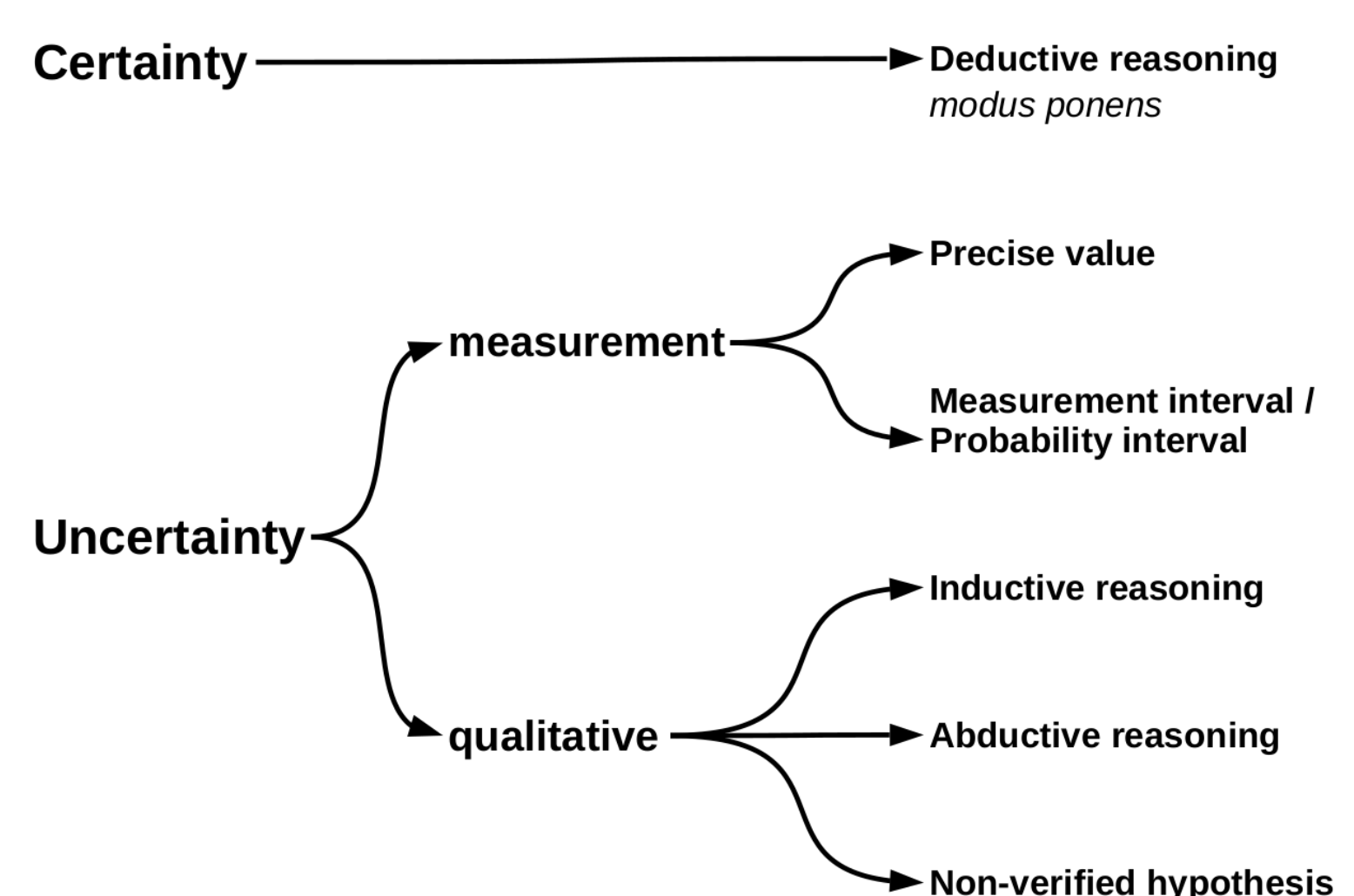
*"It is not known what these root attributes may be."*

*"Secondly, the negative effect may be that temperature rise can increase the consumption of water and bring on a water deficit in some biomes."*

*"The greater discrepancy during the day may be due to solar heating of the metal screen in which the Ta /RH probe is housed."*

## Ontology of uncertainty [2]

- The creation of the ontology and the evaluation of the overall annotation scheme used a dataset of papers related to climate change retrieved from ISTEEX.
- We have designed a formal grammar and linguistic rules to populate the ontology.
- These rules were used to produce, semi-automatically, a Gold Standard dataset of 700 annotated sentences.
- The evaluation of the rule-based annotation scheme obtained a Precision of 0.90.



SciNLP 2021 - 8 October 2021  
2<sup>nd</sup> Workshop on Natural Language  
Processing for Scientific Text