Annotation of scientific uncertainty using linguistic patterns

Panggih Kusuma Ningrum, Iana Atanassova

Published in Scientometrics (2024)
https://doi.org/10.1007/s11192-024-05009-z

Annotation of Scientific Uncertainty Using Linguistic Patterns

Panggih Kusuma Ningrum[1]*† and Iana Atanassova[12]†

[1]* Université de Franche-Comté, CRIT, F-25000 Besançon, France
[2]Institut Universitaire de France (IUF), France.

*Corresponding author(s). E-mail(s): panggih_kusuma.ningrum@univ-fcomte.fr;
Contributing authors: iana.atanassova@univ-fcomte.fr;
†These authors contributed equally to this work.

Abstract

Scientific uncertainty is an integral part of the research process and inherent to the construction of new knowledge. In this paper, we investigate the ways in which uncertainty is expressed in articles and propose a new interdisciplinary annotation framework to categorize sentences containing uncertainty expressions along five dimensions. We propose a method for the automatic annotation of sentences based on linguistic patterns for identifying the expressions of scientific uncertainty that have been derived from a corpus study. We processed a corpus of 5,956 articles from 22 journals in three different discipline groups, which were annotated using our automatic annotation method. We evaluate our annotation method and study the distribution of uncertainty expressions across the different journals and categories. The results show a predominant concentration of the distribution of the scientific uncertainty expressions in the Results and Discussion section (71.4%), followed by 12.5% of expressions in the Background section, and the largest proportion of uncertainty expressions, approximately 70.3%, are formed as author(s) statements. Our research contributes methodological advances and insights into the diverse manifestations of scientific uncertainty across disciplinary domains and provides a basis for ongoing exploration and refinement of the understanding of scientific uncertainty communication.

Keywords: scientific uncertainty, semantic annotation, annotation framework, linguistic patterns

1. Introduction

Uncertainty is an important component of scientific discovery and an integral part of the research process. The production of new knowledge uses rigorous methodological approaches based on the object of study and its disciplinary field. However, the use of tools or observations that have a margin of error, as well as the use of abductive and inductive reasoning in science, implies the presence of uncertainty. Scientists face uncertainty at different stages of the research process, from developing research questions to choosing research methods, interpreting their results, and presenting their findings to others (Cordner and Brown, 2013). Furthermore, uncertainty plays an important role in the construction of new knowledge in the experimental sciences, where the hypothetico-deductive model implies the formulation of hypotheses that need to be verified. The perception of uncertainty in scientific discourse is therefore an important issue for all scientific activity.

This research proposes an interdisciplinary annotation framework to identify and categorise sentences that express uncertainty in articles. We use this annotation framework to study a corpus of 6 journals from three different disciplines. The main objective of this study is to propose a method for annotating scientific uncertainty in texts and elucidate the diverse forms of uncertainty prevalent across disciplines, along with their contextual roles within scientific discourse. To achieve this objective, we compile a dataset of articles, analyzing two distinct samples of sentences: one derived through uncertainty cue mapping and the other through manual annotation of randomly selected articles. Additionally, our secondary objective involves constructing an annotated dataset for the development of automated tools (Ningrum and Atanassova, 2023). To achieve this, we design sets of linguistic patterns tailored for uncertainty annotation and rigorously test them on extensive corpora comprising 22 journals. These journals are distributed across three distinct academic realms: Medicine; Biochemistry, Genetics, & Molecular Biology; and a corpus that encompasses both broad-scope and interdisciplinary research. These corpora contain a substantial amount of textual data, comprising 5,956 articles and a total of 1,106,268 sentences.

This paper is organized as follows: The next section presents a literature review of relevant research on the classification and identification of scientific uncertainty in papers. The following section describes the methodology, including dataset selection and an introduction to the annotation framework, and outlines the research pipeline for two experiments: 1) Manual annotation, involving Uncertainty Cue Mapping and Manual Uncertainty Expression Search, and 2) Automatic annotation. The results section provides a detailed analysis of the frequencies and distributions of uncertainty expressions across different categories and journals for both experiments. Finally, a discussion of these results is presented.

1.1. Background

Uncertainty is a complex concept with multiple definitions (Walker et al., 2003; Refsgaard et al., 2007; Ascough Ii et al., 2008). Consequently, the literature offers a broad range of meanings and interpretations of the term. Numerous studies have used a range of techniques to identify and explore scientific uncertainty, from conducting observations using content analysis (Light, Ying Qiu, & Srinivasan, 2004; Pinto, Osório, & Martins, 2014) to more sophisticated and automated processes based on computational methods (Medlock & Briscoe, 2007).

Studies on the identification of text segments expressing uncertainty have been proposed by Atanassova, Rey, and Bertin (2018), who use the corpus of hedge verbs proposed by Hyland (1998) and an extended vocabulary of hedging and uncertainty cues proposed by Chen, Song, and Heo (2018) to generate a list of strong indicators of uncertainty and observe their distribution in articles in biomedicine and physics. In addition, Rey, Bertin, and Atanassova (2018) address the problem of interdisciplinary and conceptual understanding of the concept of uncertainty by studying a corpus of scientific articles on global warming. This work has produced a relational scheme of scientific uncertainty in which the uncertainties expressed in the texts are organised into classes according to the type of reasoning used (abductive, inductive, deductive) and the presence or absence of quantitative references to the uncertainty.

Journal articles have been found to be an ideal source for learning and exploring scientific uncertainty. The plausible reason for this is that journal articles are considered to be more detailed and reliable sources than other types of text, even when compared with other scientific writing such as technical, clinical or laboratory reports. This is because other scientific writing is rarely subjected to extensive independent peer review and is intended for internal audiences within a particular organisation. In addition, journal articles are a common medium used by scientists to communicate their structural thinking and findings to their colleagues and the scientific community. Most importantly, journal articles play an important role in disseminating knowledge to a wider audience. Journal articles are a socially situated activity through which authors connect with their audience. They not only describe the structural thinking of the author(s), depict the author's persona, and explain the research and analysis process (Candlin & Hyland, 2014; Hyland, 1996; Candlin, 2000; Hyland, 2000).

Identifying and measuring the degree of uncertainty associated with scientific knowledge in the vast and rapidly growing volume of journal articles remains a challenge (Chen, Song, & Heo, 2018). The fundamental problem is working with unstructured textual data in scientific literature. This is mainly because natural language is inherently flexible, allowing for a wide range of expressions and meanings, which complicates the process of interpreting and analysing data. Previous studies have primarily focused on detecting and identifying a specific set of hedging and uncertainty cues or markers in scientific abstracts (Vincze et al., 2008; Guillaume et al., 2017) or full-text articles (Hyland, 1996; Medlock and Briscoe, 2007; Yao, Wei, and Wang, 2023). These studies have expanded the vocabulary and lexicon of uncertainty, but implementing the technique is challenging due to the extensive manual work required and the high complexity of natural language.

1.2. Annotation framework for scientific uncertainty

As shown earlier, there exist a number of concepts and terminologies associated with scientific uncertainty, many of which are broad and general. Previous research predominantly focused on particular aspects of scientific uncertainty, such as modality, hedging, negation, or the occurrence of uncertainty cues. In addition, several typologies and ontologies of uncertainty have been developed for different purposes, some of which are domain-specific, such as an ontology of scientific uncertainty presented by Blanchemanche et al. (2013) for food risk assessment, and a typology of analytical uncertainty for geospatial information by Thomson et al. (2005). Furthermore, most of the existing approaches to identify and categorise uncertainty take into account only a single dimension of uncertainty. For instance, Budescu and Wallsten (1995) focused on linguistic representations of uncertainty, including verbal and numerical representations, while Fox and Ulkumen (2011) emphasised the

nature of uncertainty, namely epistemic and aleatory. While these approaches are useful for investigating specific domains and areas, the diverse concepts, and classifications of uncertainty in science suggest that it is a highly complex phenomenon that cannot be adequately captured by a one-dimensional framework.

The work of Walker et al. (2003) is an example of a multidimensional framework. They harmonised and integrated previous research on uncertainty (e.g., Funtowicz and Ravetz, 1990; Morgan and Henrion, 1992; Van Asselt, 2000; Van Der Sluijs, 1997) into a single coherent taxonomy for uncertainty classification. The research focused on the analysis of scientific uncertainty in model-based decision support by developing a framework and a common vocabulary for classifying uncertainty in a model. This approach represents scientific uncertainty according to three principal dimensions, i.e., location, level, and nature. The first dimension is location, which refers to where the uncertainty exists in a scientific model, such as in the system boundaries or in the model parameters. The second dimension is the level of uncertainty, which ranges from simple statistical uncertainty to total ignorance. The third dimension is the nature of uncertainty, which can arise from a lack of knowledge (epistemic uncertainty) or from the inherent variability of a phenomenon (aleatory uncertainty). This framework has been utilised by a variety of researchers who have incorporated it into their own frameworks for uncertainty analysis. For example, Meijer et al. (2006) modified this original framework to categorise perceived uncertainties in socio-technical transformations by changing the location dimension and redefined the framework to study perceived uncertainties. Fijnvandraat (2008) modified this framework to better understand the role of uncertainty and risk in infrastructure investment with a focus on broadband deployment by replacing the scale used to describe the level of uncertainty with a different one introduced by Courtney (2001).

In the field of NLP, Rubin et al. (2006) proposed a multidimensional theoretical framework for the manual categorisation of explicit certainty information in newspaper articles. This multidimensional framework has been designed considering various problems in the field of NLP, making it compatible for implementation. The certainty markers in this study are classified into four dimensions: level of certainty, perspective, focus and timeline.

However, the above-mentioned frameworks are not fully applicable to the current study. The first framework from Walker et al. (2003) is primarily concerned with model-based decision making, whereas the current study is concerned with the end-to-end research process. Furthermore, the scope of the present study includes scientific uncertainty, which is expressed in journal articles, whereas Walker's framework includes external factors, such as stackholders in the decision making process and the economic, political, social situation. The latter form of framework (Rubin et al., 2006) seems promising for the current study, as it was specifically built using NLP concepts. However, the framework focuses mainly on the identification of certainty expressions in text instead of the uncertainty expressions and its scope is limited to the manual categorisation of explicit certainty in newspaper articles, resulting in some attributes that are incompatible with the characteristics of scientific article data and the scope of the current study.

2. Methodology

Based on the concepts present in the studies described above, we present the first annotation framework of scientific uncertainty expressed in articles across different dimensions. This framework is intended to be interdisciplinary. An uncertainty categorisation model with five dimensions: Reference, Nature, Context, Timeline and Expression, is proposed. Fig. 1 shows these five dimensions and how each dimension is divided into categories. A detailed description of the dimensions is given in the following sections.

Reference

According to Stocking and Holstein (1993), a typical scientific text may contain a variety of statements and information discussing not only the current study but also previous studies. This theory serves as the foundation for the first dimension in the current framework, which addresses the 'who' or reference of the expression of scientific uncertainty, whether it refers to the author(s) of the observed journal article or to the third party or author(s) of previous research.

The last group of this category, "Both author(s) & former study(ies)", is intended to accommodate complex sentences that may refer to both the author(s) and the previous study(s). To accurately classify a sentence under this category, it is necessary to examine closely the argument structure of the sentence, the manner in which

previous studies are cited, and the positional context of the citations. The aim is to ensure clarity and avoid misinterpretation that could arise from the complex interplay of multiple phrases within a single sentence.

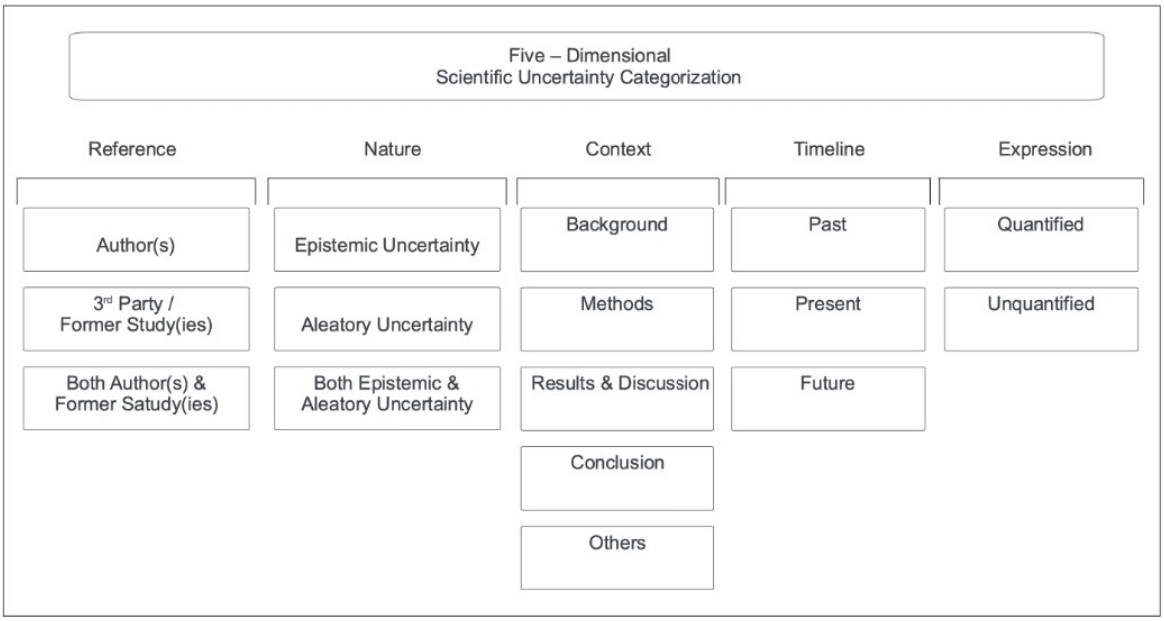


Fig. 1 Framework for Scientific Uncertainty Categorization

Nature

The second dimension of uncertainty is whether the uncertainties are caused by a lack of knowledge (epistemic) or by inherent variability (aleatory) in the system itself. Assessing the nature of the uncertainty can help to understand how specific uncertainties can be addressed. This dimension can be divided into two categories:

Epistemic uncertainty. Epistemic uncertainty refers to deficiencies caused by a lack of knowledge or the complexity of information. In theory, knowledge creation and learning can help to reduce this type of uncertainty. Other terms for epistemic uncertainty include knowledge, internal, secondary, or substantive uncertainty (Meijer et al., 2006; Dosi and Egidi 1991; Helton, 1994; Jauch and Kraft, 1986; Kahneman and Tversky, 1982; Rammel and Bergh, 2003; Van Asselt and Rotmans, 2000; van der Sluijs, 1997; Walker et al., 2003).

Aleatory Uncertainty. Aleatory uncertainty refers to the uncertainty arising from inherent variability or uncertainty introduced by probabilistic variations in a random event. Although aleatory uncertainty cannot be eliminated, it can be managed by determining the relative propensities of events. Other terms for aleatory uncertainty are variability, strong, fundamental, stochastic, random, primary, external, procedural, or ontological uncertainty (Dosi and Egidi 1991; Helton, 1994; Jauch and Kraft, 1986; Kahneman and Tversky, 1982; Rammel and Bergh, 2003; Van Asselt and Rotmans, 2000; van der Sluijs, 1997; Witteloostuijn, 1986; Walker et al., 2003).

Additionally, the last group in this category, “Both Epistemic & Aleatory”, is for complex sentences that may express both types of uncertainty.

Context

The context of uncertainty is the way in which the uncertainty itself appears in the journal article. According to Friedman and Kandel (1999), each section of a scientific text may contain varying degrees of uncertainty. The current study uses this logic as the basis for the third dimension of the framework, as journal articles typically use the IMRaD format: Introduction, which basically represents the background and rationale of the study, Methodology, Results and Discussion, Conclusion, and Other.

Timeline

The fourth dimension considers the relevance of time (past, present, and future) to the moment the article is written. The past naturally includes completed or recent states or events; the present includes current, immediate, and incomplete conditions; and the future includes predictions, plans, warnings, and proposed actions. This dimension is based on the work of Rubin et al. (2006).

The Timeline dimesion does not include labels that combine uncertainties related to Past and Present or Present and Future. This is due to the fact that the analysis of our dataset showed that the vast majority of uncertainties can be labeled with only one of the labels of the Timeline dimension.

Expression

The final dimension is concerned with how uncertainty is presented and communicated in the text. This dimension is divided into two subcategories:

Quantified. Quantifiable uncertainty can be expressed in absolute quantitative terms, including a probability distribution or confidence interval, or in relative terms, such as likelihood ratios, or in an approximate quantitative form, verbal summary, and so on. Other terms for quantifiable uncertainty include first-order uncertainty and direct uncertainty (Bles et al., 2018).

Unquantified. Unquantified uncertainty can be expressed as a set of caveats about the underlying sources of evidence, which can be combined into a qualitative or ordered categorical scale. Second-order or indirect uncertainty are other terms for unquantified uncertainty.

Table 1 presents some examples of sentences with their annotations using the above categories.

Table 1 Examples of sentences and annotations

Sentence	Journal	Uncer-tainty	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	Yes	Former / Previous Study(s)	Epistemic	Background	Past	Unquantified
It is possible that corticosteroids prevent some acute gastrointestinal complications.	BMC Med	Yes	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	Yes	Author(s)	Epistemic	Results & Discussion	Future	Unquantified
In this test, a likelihood ratio test statistic is calculated for the 'two tree' versus 'one tree' models, and compared to a null distribution generated by non-parametric bootstrapping (see Methods).	PLoS One	No	-	-	-	-	-
This may be due to the increased prevalence of short-term relationships and regularly changing sexual partners [74], which, as a result, might lead to less chance of starting a family.	BMC Med	Yes	Both author(s) & former/previous study(s)	Both aleatory & epistemic	Results & Discussion	Present	Unquantified

2.1. Dataset

In the present study, the pre-defined criteria used to select scientific articles for the dataset included (1) peer-reviewed articles from high-quality and reputable international journals, (2) written in English, (3) open access, and (4) formatted in HTML, XML, or JSON.

The first criterion acts as a primary filter, allowing the selection of high-quality data for the construction of corpora. To this end, the data in this study are derived from journals indexed in three high-quality and popular indexing databases, namely PubMed, Scopus, and Web of Science. PubMed is a well-known database that primarily covers journal literature in the biomedical and life sciences, while Scopus and Web of Science cover most scientific fields. The Scimago Journal & Country Rank (SJR) indicator is also taken into account when selecting journals, as higher SJR indicator scores are expected to indicate higher journal prestige due to its rigorous system for evaluating and analysing scientific topics. By passing this criterion, the journal articles have established a sufficiently authoritative position in the subject areas and have demonstrated noteworthy academic quality.

The second selection criterion is that the articles must be published in English, as the majority of international journal articles are written in English. The articles collected in the current study could have been written by non-native English speakers, but they are still included in the corpus because scholarly articles published in prestigious journals and trusted worldwide databases are expected to follow standard English.

Articles must also meet the third condition: open access. The term ”open access” refers to the ability to access and download scholarly works free of charge. This is necessary in order for the data collected to be copyright-free for distribution via corpora.

The fourth data selection criterion is that the text data be formatted in HTML, XML, or JSON. This criterion is significant because the current study will rely on the entire text of the articles as its primary source of information. Collecting text data in HTML, JSON, and XML formats is more manageable because it eliminates the possibility of damaged text during the corpora construction procedure.

It is important to note that the nature of the research and the use of a particular word in one field may be different from that in another field. To better reflect the range of scientific inquiry, we have established three distinct corpora, sourced from journals that specialize in (1) Medicine, (2) Biochemistry, Genetics, and Molecular Biology, and (3) a corpus that encompasses both broad-scope and interdisciplinary research. The intent is to scrutinize the various ways uncertainty is expressed across different scientific domains.

The Scimago Journal & Country Rank (SJR) classification was chosen to classify and select the journals in each corpus, as it includes journal and country scientific indicators developed from information contained in Scopus, the world’s largest database of academic literature. Firstly, the journals from the SJR ranking list were filtered and selected on the basis of the category labels assigned. Journals that appeared in more than one subject area were excluded from the list, as each group was intended to present data that reflected the uniqueness and disciplinary purity of its subject area. Furthermore, the top two journals were selected for the Medicine; and Biochemistry, Genetics & Molecular Biology corpus.

Additionally, to incorporate the breadth of scientific study and the interplay between various fields, PLoS One and Nature were included in our comprehensive corpora. These journals were selected for their role in publishing research across a wide spectrum of disciplines, as well as studies that cross these boundaries. Both journals rank in the first quartile for a broad academic scope within Scopus and Web of Science and host extensive collections of research articles that meet the data selection criteria.

After obtaining the list of journals, the data harvesting procedure was carried out in Python and Google Cloud. First, metadata was retrieved from the Elsevier API using the elsapy module with journal names and ISSNs as input. The metadata information was then used to retrieve the full text data.

This study would only focus on the article type data. Therefore, other types of data such as Editorial, Correction, Commentary, Corrigendum, Erratum, etc. were omitted. After that, the data were saved and prepared for the data cleansing and data pre-processing phase.

Data cleansing was performed by removing irrelevant elements such as tables, figures, boxed text, graphs, supplementary material, formulas, and quotations, leaving only the clean text in each article. The text was then parsed based on its format and divided into groups containing metadata, sections, paragraphs, and sentences. The sections, paragraphs and sentences were then stored in a MySQL database.

2.2. Research Pipeline

Seven main stages were employed to achieve the objectives of the present study. They are: (1) Uncertainty Cues Lexicon (UCL) construction, (2) Data Sampling, (3) Uncertainty cues mapping process, (4) Manual Uncertainty Expression Searching process, (5) Manual Annotation, (6) Construction of linguistic patterns for annotation, and (7) Automatic Annotation. Three inputs are used: Lists of uncertainty cues and markers from Hyland (1996), Chen, Song, and Eun Heo (2017), and Bongelli et al. (2019); scientific articles that are stored on a MySQL database; and the Five-Dimensional Scientific Uncertainty Categorization. **Fig. 2** describes the stages involved in this study.

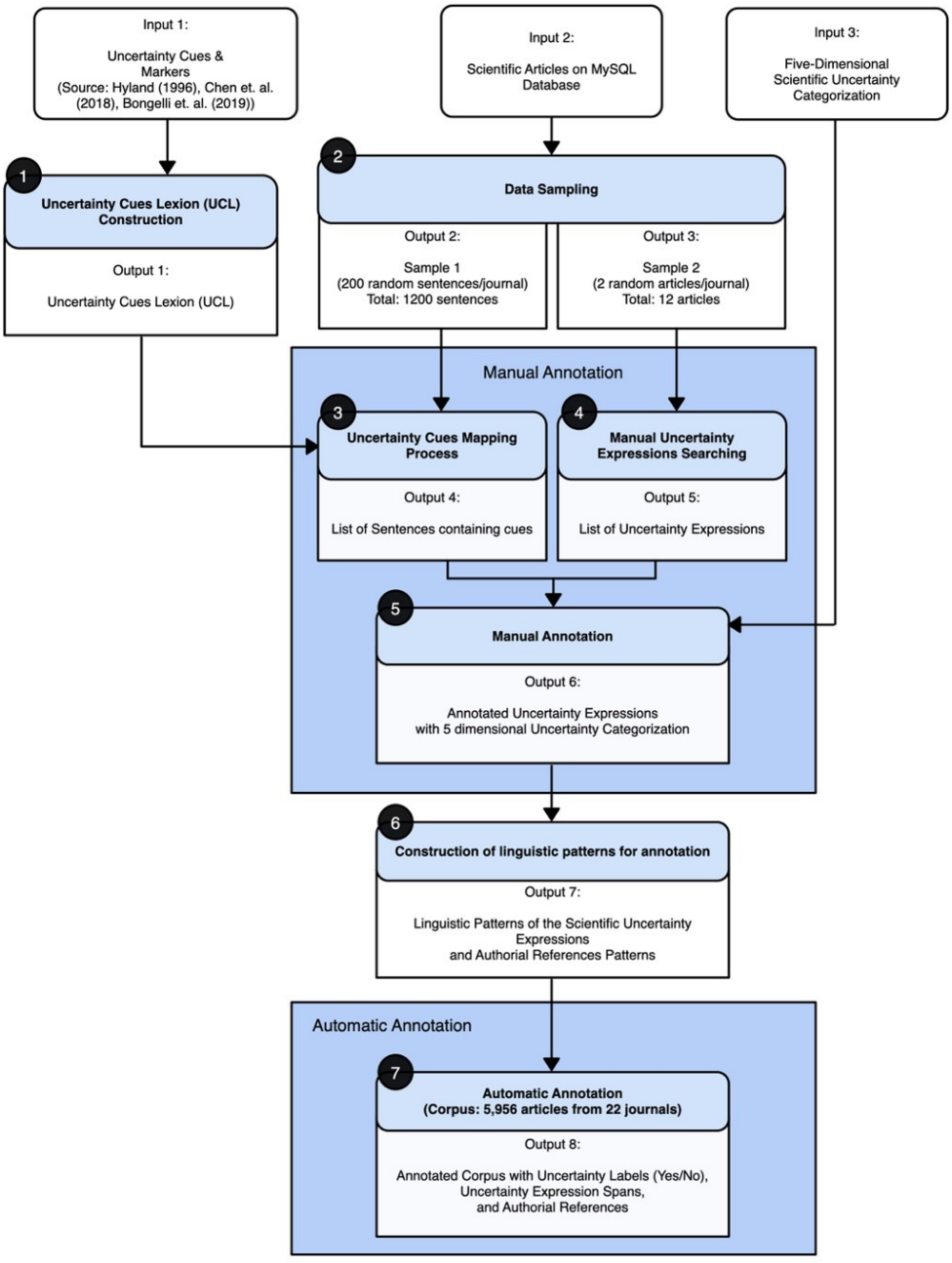


Fig. 2 Research Path Diagram

1. Uncertainty Cues Lexicon (UCL) construction

The identification of cues and markers serving as indicators for the occurrence of uncertainty expressions in texts has been a focal point in several studies (Atanassova et al., 2018; Hyland 1996; Chen, Song, and Eun Heo 2017; Bongelli et al. 2019), forming the foundational premise for our investigation. In our study, as illustrated in **Table 2,** we draw upon the insights of Hyland (1996), Chen, Song, and Eun Heo (2017), and Bongelli et al. (2019) . From these sources, a list of uncertainty cues and markers is adopted and amalgamated to construct the Uncertainty Cues Lexicon (UCL). The UCL will play a crucial role as one of the inputs in our research, providing a foundation for the analysis of uncertainty expressions in the examined texts.

Table 2 List of cues that compose the Uncertainty Cues Lexicon (UCL)

Source	Category	Cues / markers
Hyland (1996a)	-	would (not); may (not); could; might (not); should; cannot; will (not); must; shall; ought to
Chen et al., (2018)	-	unclear; suspect; controversial; ambiguity; inconclusive; unexpected; consensus; contrary; inconsistent; paradoxical; confusing; unusual; uncertain; flaw; uncertainty; dispute; unknown; impossible; ambiguous; misleading; incomplete; unexplained; contradictory; contentious; paradox; incompatible; surprising
Bongelli et. al. (2019a & 2019b)	Epistemic verbs	I suppose; suggest/s; seem/s; suggesting; assuming; we think; I believe; it seems; expect; appear; look/s; suspect/suspected; do not seem; no one has proven; not sure
	Epistemic non-verbs (Adjectives, Adverbs, Nouns, Personal Attributions)	unlikely; likely/morelikely; probably; perhaps; maybe; possible; possibility; seemingly; likelihood; not likely; plausibility; possibly; potentially; potential; to our knowledge; unclear; according to my view; in my opinion; perhaps; doubt; impression; probably; unclear; apparently; uncertainty; uncertain; apparent; assumption; confident; hypothesis; plausibly
	Modal verbs in the simple present	can; may; may not; must
	Modal verbs in the conditional mood	could; would; might; should

2. Data Sampling

In pursuit of the research objectives, namely the construction of linguistic patterns and the identification of an effective automatic annotation method, two data sampling methods were employed. The initial approach involved the random selection of 200 sentences from the MySQL database corresponding to each target journal. This initial sample, denoted as Sample 1, comprised a total of 1200 sentences, serving as the input for the uncertainty cue mapping process.

Simultaneously, a second sampling method was employed, wherein two articles were randomly chosen from each targeted journal. This subsequent sample, denoted as Sample 2, consisted of 12 articles and was specifically utilized for the manual identification of uncertainty terms. This manual search is pivotal in accommodating the inherent flexibility of natural languages, recognizing that uncertainty expressions may exhibit diverse linguistic patterns and variations that do not strictly adhere to predetermined cues or markers. Hence, this step was incorporated to ensure the inclusivity of potential variations in expressions, enhancing the robustness of our analysis.

3. Uncertainty Cues Mapping Process

In this step, Sample 1, consisting of 200 randomly selected sentences from the MySQL database for each target journal, and the Uncertainty Cues Lexicon (UCL) served as inputs. Taking advantage of the efficiency of regular expressions (RegEx), we carried out the process of mapping the uncertainty cues. Practically, each cue within the UCL was systematically mapped to Sample 1, followed by identifying and compiling all sentences containing such cues. This systematic approach allowed for a comprehensive analysis, efficiently extracting sentences indicative of uncertainty for further investigation and annotation.

4. Manual Uncertainty Expressions Searching Process

This step involved a meticulous manual search using Sample 2 as dataset. Each sentence within each selected article was subjected to a comprehensive screening process, with a strong focus on the identification and marking

of sentences that conveyed expressions of uncertainty. Furthermore, the marked sentences were meticulously compiled into a coherent list, strategically organised and prepared for the subsequent annotation process.

5. Manual Annotation process

Two annotators were involved in this process. The outputs of the Uncertainty Cues Mapping Process and the Manual Uncertainty Expressions Searching Process were used as data. The annotation process used the Five-Dimension Scientific Uncertainty Categorization, and each sentence was annotated as either containing ”Uncertainty” or “No Uncertainty” and then annotated with the categories of the five dimensions.

Each annotator was provided with a set of explicit instructions that included guidelines for the annotation process. Additionally, a collection of previously annotated text data was provided as a reference. In order to ensure the accuracy and consistency of the annotations, both annotators underwent training and testing in which they labelled the data jointly. This practice facilitated discussion between the annotators and ensured the development of a coherent understanding of the guidelines and labelling standards. Then, the two annotators worked independently to label the dataset. Upon completion of the annotation process, any inconsistencies were resolved through discussion and consensus. In very rare cases where the annotators could not agree on a particular label, a third annotator was called in to make a final decision.

6. Construction of linguistic patterns for annotation

Based on the annotated dataset that was constructed in the previous steps, we observed the linguistic patterns present in the sentences that express uncertainty. A linguistic pattern is composed of one or more keywords (cues) and sets of linguistic elements that occur in their contexts in the same sentence. Thus, these patterns are complex structures that include uncertainty cues, but also allow to eliminate most of the ambiguity and noise in the annotation compared to the use of simple cue words.

We identified sets of patterns that fulfill the following criteria:

- the presence of a linguistic pattern in a sentence implies that the sentence expresses uncertainty and therefore the sentence should be annotated;
- the elements of a pattern can be useful to identify some of the categories of uncertainty, i.e. annotate the sentence with these categories, according to our annotation framework.

A detailed description of this process can be found in the seminal work by Ningrum, Mayr, and Atanassova (2023).

7. Automatic annotation

We have devised and implemented a sophisticated system designed to identify and annotate sentences expressing uncertainty within scientific texts, leveraging the linguistically defined patterns described earlier. The operational pipeline of this system incorporates a multifaceted approach, encompassing pattern matching, scrutiny of complex sentence structures, and annotation of authorial references. These systematic steps collectively enable the accurate identification and annotation of expressions of scientific uncertainty (Ningrum, Mayr, and Atanassova, 2023).

To assess the efficacy of our automatic annotation system, we conducted an initial evaluation using the annotated corpus manually constructed in the preceding steps. The system demonstrated commendable performance metrics, achieving an accuracy of 0.898, a precision of 0.928, a recall of 0.920, and an F1 Score of 0.924. These results underscore the system’s reliability, particularly considering the scale of this initial dataset.

Subsequently, we applied our system to annotate a larger dataset comprising 22 journals and 5,956 articles. In the following section, we present in detail the results of this automatic annotation process.

3. Results

In this section, we present the results of scientific uncertainty identification and categorisation from two experimental settings, namely: 1) Manual annotation, involving Uncertainty Cue Mapping and Manual Uncertainty Expression Search, and 2) Automatic annotation.

3.1. Manual Annotation

Table 3 depicts the results of cue mapping on the total sample of 1200 sentences. For each discipline and journal, we show the number of articles and sentences that contain cues, the number of cues that correctly represent expressions of uncertainty, and their percentage within the journals.

Table 3 Results of cue mapping on the total sample of 1200 sentences

Discipline	Journal	Tot. Articles with cue(s)	Tot. Sentences with cue(s)	Cue occurrences (n)*	% of Sentence with cue by tot. samples in journal	Uncertainty Occurrances (sentences)	% of Uncertainty Occurrances by tot. samples in journal	% of Uncertainty Occurrances by total cues
Medicine	BMC Med	49	58	84	29.00%	32	16.00%	38.10%
	Cell Mol Gastroentero l Hepatol	23	31	37	15.50%	8	4.00%	21.62%
Biochemistry, Genetics & Molecular Biology	Nucleic Acids Res	50	50	60	25.00%	21	10.50%	35.00%
	Cell Rep Med	20	29	37	14.50%	12	6.00%	32.43%
Interdisciplinary & miscellaneous	Nature	32	43	61	21.50%	18	9.00%	29.51%
	PLoS One	40	47	65	23.50%	16	8.00%	24.62%
Total		214	258	344	21.50%	107	8.92%	31.10%

**more than one cue can occur in one sentence*
Total sentences randomly selected as samples: 1,200 (200/Journals)

The results of the uncertainty cue mapping process indicate that in the total sample of 1200 sentences, 258 sentences (21.50%) were identified as containing uncertainty cues. Of these, up to 107 sentences (8.92%) were annotated as expressing uncertainty. Among the journals, BMC Medicine (32) contributes the highest number of sentences with uncertainty in the dataset, followed by Nucleic Acids Research (21), Nature (18), PloS One (16), Cell Reports Medicine (12), and Cellular and Molecular Gastroenterology and Hepatology (8).

Additionally, we observe that only about 31% of the sentences containing cues express uncertainty. This means that the cues in the UCL list can only be considered as weak indicators of uncertainty and their presence alone is not sufficient to annotate the corpora. The majority of sentences containing cues were discarded by the human annotators as not expressing uncertainty. Examples of such sentences are:

- "*With these vectors, anti-cancer drugs can be delivered to tumors much more effectively than by circulatory delivery alone [23].*" (BMC Med)
- "*Because of the rapidity with which we could obtain these cells, we could implant them into aneuronal muscle explants from the same individual.*" (Cell Mol Gastroenterol Hepatol)
- "*A form of antenatal education needs to be delivered which gives expectant mothers a more realistic expectation of what is likely to happen in labour [37].*" (BMC Med).

Fig. 3 shows the frequency of occurrence of all uncertainty cues. Overall, the modal verbs and the cues from the list of (Hyland 1996) tend to be more frequent than the epistemic non-verbs. At the same time, we know that modal verbs are particularly polysemic, which means that their presence in sentences can be associated with a variety of meanings that are not necessarily related to uncertainty. Among all cues, the five most frequent uncertainty cues occurring in the dataset are 'may' (47), 'when' (27), 'can' (26), 'could' (25) and 'if' (19). Furthermore, the results show that a sentence can contain more than one uncertainty cue. Among the 153 sentences containing multiple cues, we found that there are up to 95 sentences (62.1%) that express uncertainty.

Additionally, the results of the uncertainty manual searching process revealed that in the sample of 12 articles, a total of 95 sentences were annotated with occurrences of uncertainty. **Table 4** presents their distribution in the

different journals. The number of sentences in each journal varies from 5 to 36. This may be due to the small size of this sample, as only two articles per journal were examined.

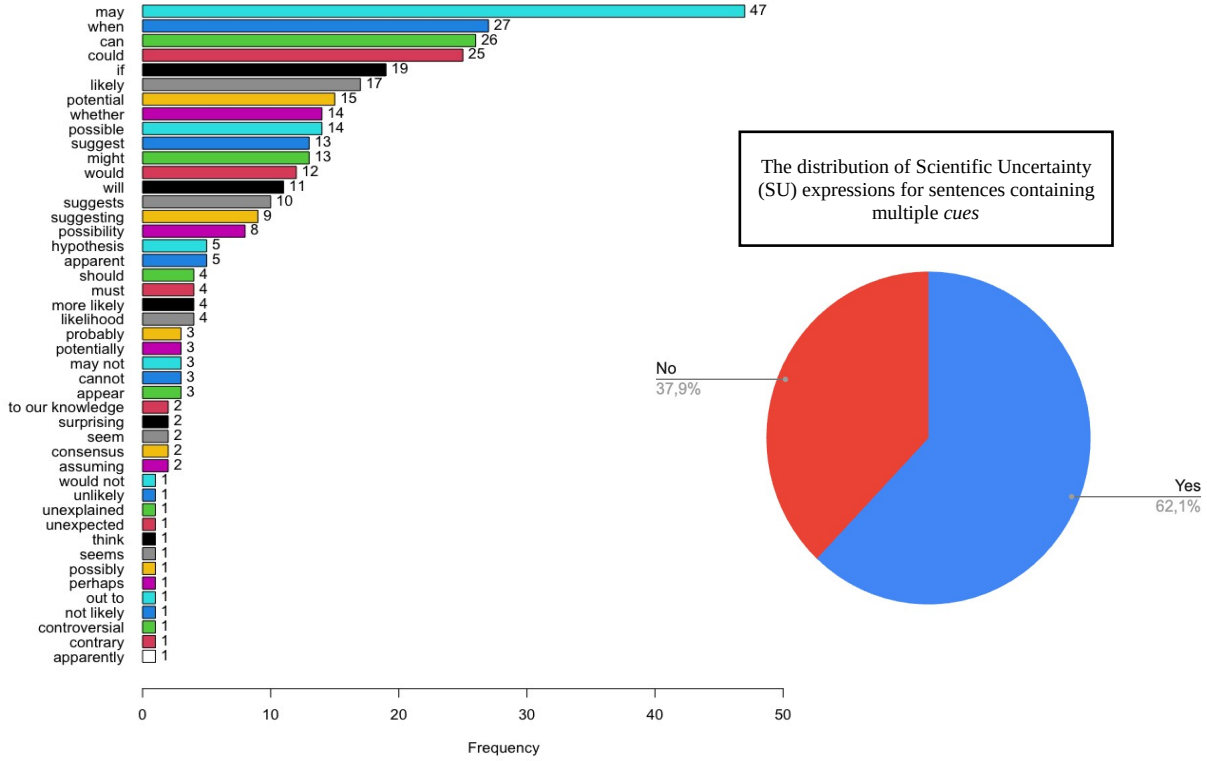


Fig. 3 Uncertainty Cues Occurrences and The Distribution of Scientific Uncertainty (SU) Expressions for Sentences Containing Multiple Cues

Table 4 Results of manual searching on the total sample of 12 articles

Discipline	Journal	Uncertainty Occurrances in sentence (unique n)
Medicine	BMC Med	36
	Cell Mol Gastroenterol Hepatol	5
Biochemistry, Genetics & Molecular Biology	Nucleic Acids Res	13
	Cell Rep Med	19
Interdisciplinary & miscellaneous	Nature	14
	PLoS One	8
Total:		**95**

Total articles randomly selected as Samples: 12 (2/Journals)

Table 5 provides a comprehensive overview of the distribution of sentences expressing uncertainty across different categories. The distributions resulting from both cue mapping and manual searching methods exhibit notable similarities.

In the cue mapping process, significant disparities emerge across categories. In the Reference dimension, the preeminent annotation for uncertainties (87.0%) is attributed to "Author(s)," while "Previous studies" and the combination of both account for 10.5% and 2.5%, respectively. In terms of the Nature, Epistemic uncertainty predominates, constituting 77.2%, contrasting with Aleatory uncertainty at 22.8%. Concerning the Context dimension, a substantial majority of uncertainties (57.04%) manifest in the Results and Discussion section, with the remaining sections contributing less significantly. The "Others" section follows closely with 21.6%, trailed by the Background section at 17.9%. The Timeline dimension reveals that Past and Future categories collectively

account for less than 20%, with the Present dominating at 81.5%. Notably, the overwhelming majority of uncertainties are characterized as Unquantified, with a mere 0.6% classified as Quantified.

Table 5 Uncertainty distribution by categories (Cue Mapping & Manual Searching Results)

Uncertainty Category		Cue Mapping (Total: 1200 sentences)		Manual Searching (Total: 12 articles)	
		Frequency (n)	Proportion in each Category (%)	Frequency (n)	Proportion in each Category (%)
Reference	Author(s)	141	87.0%	84	88.4%
	Former Study(ies)	17	10.5%	8	8.4%
	Both	4	2.5%	3	3.2%
Nature	Epistemic	125	77.2%	77	81.1%
	Aleatory	37	22.8%	16	16.8%
	Both	0	0,00%	2	2.1%
Context	Background	29	17.9%	18	18.9%
	Method	4	2.5%	3	3.2%
	Results & Discussion	93	57.4%	44	46.3%
	Conclusion	1	0.6%	6	6.3%
	Others	35	21.6%	24	25.3%
Timeline	Past	25	15.4%	14	14.7%
	Present	132	81.5%	73	76.8
	Future	5	3.1%	8	8.4%
Expression	Quantifiable	1	0.6%	0	0%
	Unquantifiable	162	99.4%	95	100%

In the manual searching process, an overwhelming proportion (88.4%) of sentences are annotated as "Author" in the Reference category. Epistemic nature constitutes 81.1% of sentences, while Aleatory or both represent the remaining. Approximately 46% of uncertainties are concentrated in the Results and Discussion sections. In terms of the Timeline dimension, the majority is annotated as "Present" (76.8%), with a comparatively small number annotated as "Past" and "Future." Strikingly, all sentences in this sample are annotated as Unquantified, with no instances of Quantified expressions identified. These detailed findings shed light on the nuanced distribution of uncertainty expressions across diverse dimensions and categories.

3.2. Results of the Automatic Annotation

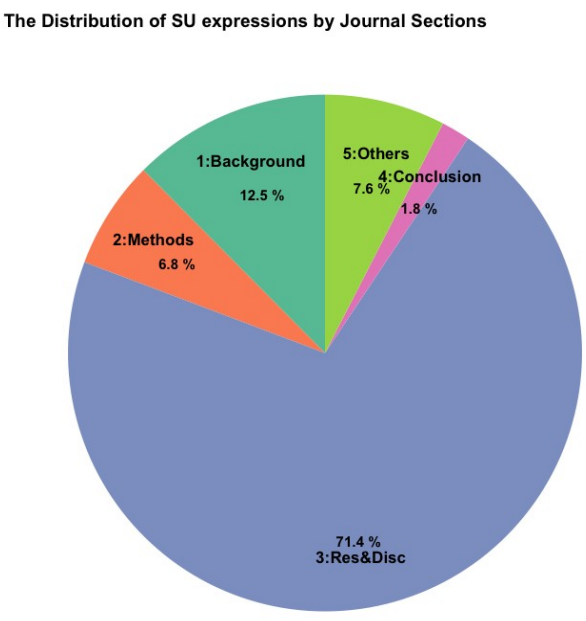
Table 6 shows the total number of sentences and articles in each journal and the distribution of scientific uncertainty expressions automatically annotated in the corpora. The analysis of the corpora containing 1,106,268 sentences from 22 journals revealed a remarkable prevalence of uncertainty expressions, representing 163,496 sentences, or 14.5% of the total dataset.

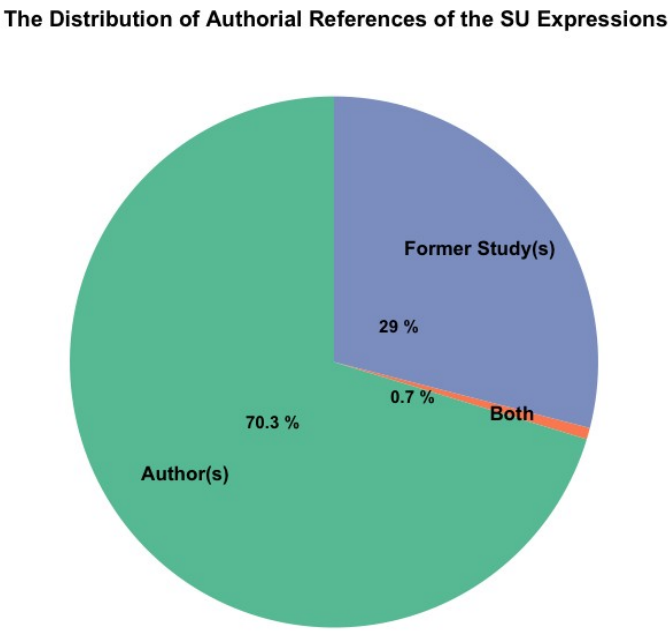
The distribution of uncertainty expressions varies significantly between the journals in our dataset. The highest frequencies are observed in European Psychiatry (23.53%), Clinical Epidemiology (18.68%) and BMC Medicine (18.39%). Conversely, the lowest frequency of uncertainty expressions is found in Cellular and Molecular Gastroenterology and Hepatology, where they represente 9.36% of the sentences analysed. The average frequency for all journals is 15.51%.

Table 6 The Distribution of Scientific Uncertainty Expressions Detected in the Corpus (Automatic Annotation)

Discipline	Journal	Total Article	Total Sentences	Uncertainty Expression

		on DB	on DB	detected	
Medicine	BMC Medicine	535	93,700	17,235	18.39%
	Cellular and Molecular Gastroenterology and Hepatology	583	176,597	16,538	9.36%
	Emerging Infectious Diseases	73	6,633	1,113	16.78%
	Cardiovascular Diabetology	92	11,276	1,728	15.32%
	Journal of Stroke	123	11,491	2,017	17.55%
	Annals of Intensive Care	54	7,392	1,235	16.71%
	BMJ Global Health	101	17,784	3,103	17.45%
	Clinical Epidemiology	123	15,160	2,832	18.68%
	Respiratory Research	94	14,827	2,075	13.99%
	European Psychiatry	15	2,678	630	23.53%
Biochemistry, Genetics & Molecular Biology	Nucleic Acids Research	1,871	312,492	51,635	16.52%
	Cell Reports Medicine	263	89,652	9,339	10.42%
	Signal Transduction and Targeted Therapy	56	11,873	1,325	11.16%
	Nature Communications	85	14,885	2,215	14.88%
	Cell Reports	173	35,591	5,459	15.34%
	Cell Discovery	155	37,676	4,938	13.11%
	EMBO Molecular Medicine	109	24,838	3,708	14.93%
	Aging Cell	86	16,389	2,688	16.40%
	Molecular Metabolism	87	18,199	2,729	15.00%
	Stem Cell Reports	124	24,646	3,481	14.12%
Interdisciplinary & miscellaneous	Nature	832	108,153	14,759	13.65%
	PLoS One	322	54,336	9,737	17.92%
Total		**5,956**	**1,106,268**	**160,519**	**14.51%**





The analysis of uncertainty expressions within the article shows a predominant concentration in some specific sections. In particular, 71.4% of uncertainty expressions were identified in the Results and Discussion section, followed by 12.5% of expressions in the Background section. It is noteworthy that the largest proportion of uncertainty expressions, approximately 70.3%, are authorial statements. A detailed breakdown of this distribution is given on **Fig. 4**. These graphs illustrate the precise location of uncertainty expressions within the article and provide a comprehensive view of their distribution.

Fig. 4 The Distribution of SU Expressions in Articles

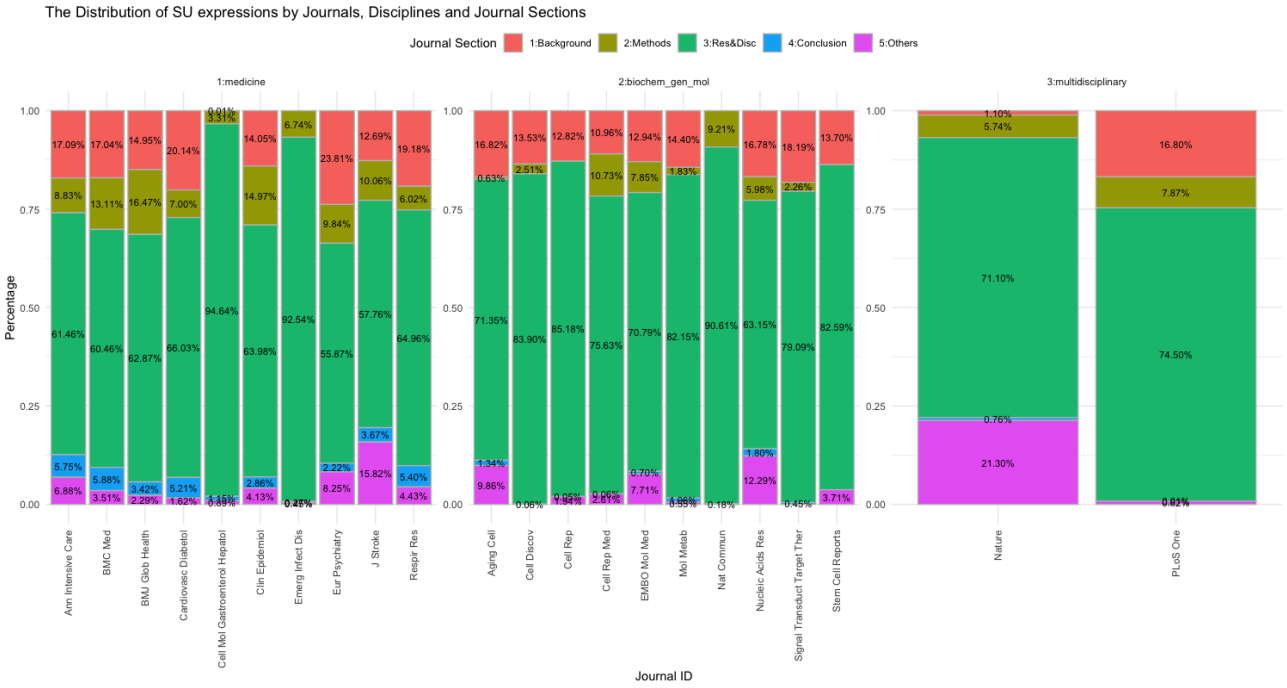


Fig. 5 The Distribution of SU Expressions by Journals, Disciplines and Journal Sections

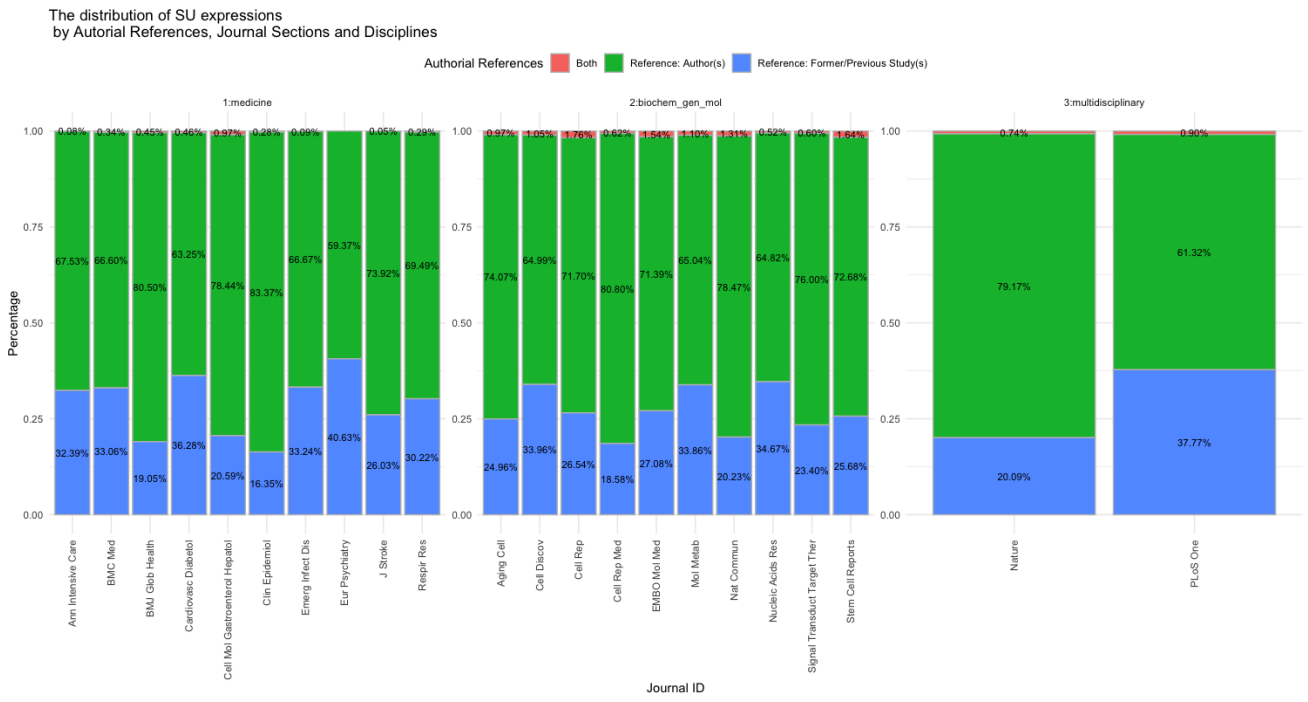


Fig. 6 The Distribution of SU Expressions by Authorial References, Journal Sections, and Disciplines

Upon further analysis of uncertainty expression distribution across journals and disciplines, a consistent trend emerged. **Fig. 5** illustrates that roughly 73.7% of uncertainty expressions across all journals were concentrated in the Results and Discussion section. Similarly, confirming a uniform prevalence across disciplines, **Fig. 6** indicates that the predominant form of detected uncertainty expressions was in the form of author(s) statements, constituting approximately 71.3% across all journals.

4. Discussion

As the notion of uncertainty is complex in nature, our study provides a first approach to characterising its multiple dimensions and observing its distributions in scientific corpora. Our corpus study is limited in several ways. First, the size of the two samples we studied is relatively small (1200 sentences and 12 articles), which may lead to over- or under-representation of some categories. We plan to conduct studies with larger samples. However, the human effort required for this kind of annotation is important, as each sentence has to be carefully examined and annotated according to five dimensions. Secondly, the selected disciplines and journals are small, only 2 disciplines and 2 journals representing both broad-scope and interdisciplinary research. This choice was partly determined by the availability and ease of harvesting of open access datasets. In the future, a wider range of scientific fields should be considered in order to observe inter-disciplinary differences in the way uncertainty is mobilised. The samples from PloS One and Nature, representing interdisciplinary and miscellaneous journals, do not have sufficient records to observe this.

The sampling methods were chosen to use existing resources (cue lists from previous studies) to select a first sample of sentences that are likely to express uncertainty. Our experiment shows that such cues are not sufficient to identify sentences expressing uncertainty. In fact, only a few of these sentences were annotated with uncertainty (about 31%). On the other hand, it is possible that a sentence expresses uncertainty but does not contain any of the cues from the list. In order to have the possibility of identifying such sentences, we constructed the second sample, which is obtained by randomly selecting articles that are fully analysed manually. We observe that the distributions on the two samples are quite similar, which can be an indication that the lists of cues are relevant for selecting candidate sentences for annotation.

The results of the automatic annotation of the dataset show a concentration of uncertainty expressions in the Results and Discussion section, as well as in the Background section of the articles. These results are consistent with several theoretical underpinnings and empirical studies.

In particular, the Results and Discussion section, which serves as a locus for interpreting empirical findings, acknowledging study limitations, and describing implications, inherently accommodates uncertainty expressions. Salager-Meyer (2017) observes an increased occurrence of hedging in the Discussion section compared to other sections of scientific articles. This is attributed to the discursive and speculative nature of the Discussion, where authors articulate controversial ideas that require protection from potential counterarguments. Notably, this observation is consistent with the findings of a study by Atanassova, Rey, and Bertin (2018), which examined the distribution of uncertainty expressions in biomedical and physics journals. In this study, it was found that the discussion section contained the majority of uncertainty expressions.

At the same time, the Background section establishes the foundation of the study by articulating previous research, knowledge gaps, and the rationale for the investigation. Swales (2014) asserts that this section, a hypothesis-generating introduction, introduces the unknown or poorly understood and refers to previous research relevant to the study. The background section, by design, encompasses the uncertainties that the research seeks to address.

Furthermore, empirical evidence, such as that presented by Bongelli et al. (2019), supports the prevalence of uncertainty expressions in these sections. The majority (69%) of uncertainty expressions in biomedical articles following the IMRAD format occurred in the Discussion section, with 11% occuring in the Introduction or Background section.

Based on the aforementioned points, our findings highlight the effectiveness of our proposed framework and annotated dataset, positioning them as robust foundations for advancing scientific uncertainty exploration. This suggests their wider applicability not only within the scope of our study, but also as valuable resources for interdisciplinary corpora, contributing to a nuanced understanding of uncertainty across scientific domains.

Our approach considers sentences as the basic units that are analysed and annotated. In many cases, complex sentences can express several different types of uncertainties that would be contained in different clauses. Thus, considering the segmentation of sentences into clauses could potentially avoid annotations with multiple labels. However, examples from our dataset indicate that in a majority of cases, considering only one clause in a sentence is not sufficient for determining the type of uncertainty and a larger context should be examined.

5. Conclusion

In this paper we have introduced an interdisciplinary five-dimensional framework for categorising uncertainty in articles. We conducted a corpus study with two experiments on samples of manually annotated sentences from different disciplines . The two samples of annotated sentences form a dataset that can be further used to automate some aspects of the annotation process. We observed the distribution of uncertainty categories across journals and disciplines.

This study of uncertainty can be extended by analysing larger corpora covering a wider range of disciplines. We will focus on this task in the future, with the aim of creating large-scale resources that can be used to implement automated annotation tools. The study of uncertainty on large corpora is important and can be used in a variety of applications, such as identifying novel and unsolved problems in a given scientific field, detecting incomplete theories or reasons for controversy.

Acknowledgments.

The authors gratefully acknowledge the French ANR for funding this research under the ANR InSciM project. Special thanks to the CRIT Laboratory at the Université de Franche-Comté for their support and the provision of conducive working facilities. Additionally, this paper represents a substantially extended version (at least 25% new material) of the ISSI2023 conference paper, and the authors appreciate the insights gained from the conference that contributed to the refinement of the research. The original conference paper entitled *"Scientific Uncertainty: an Annotation Framework and Corpus Study in Different Disciplines"* is accessible at: https://www.conftool.pro/issi2023/index.php?page=browseSessions&form_session=41.

Funding.

This work is supported by the ANR InSciM Project (2021-2025), funded by French ANR, grand number ANR-21-CE38-0003-01.

Availability of data and material.

The data supporting the paper is publicly accessible on the Zenodo platform at https://doi.org/10.5281/zenodo.8024787. Furthermore, the system used for automatic annotation in this study can be accessed for demonstration at https://bit.ly/unscientify-demo

Cite this article

Ningrum, P.K., Atanassova, I. Annotation of scientific uncertainty using linguistic patterns. Scientometrics (2024). https://doi.org/10.1007/s11192-024-05009-z

References

Ascough, J. II, Maier, H., Ravalico, J., & Strudley, M. (2008). Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological modelling, 219*, 383-399. doi: 10.1016/j.ecolmodel.2008.07.015

Atanassova, I., Rey, F., & Bertin, M. (2018). Studying Uncertainty in Science: a distributional analysis through the IMRaD structure. In *7th International Workshop on Mining Scientific Publications (WOSP 2018) at 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*. Miyazaki, Japan. http://lrec-conf.org/workshops/lrec2018/W24/pdf/10_W24.pdf

Blanchemanche, S., Rona-Tas, A., Cornuéjols, A., Duroy, A., & Martin, C. (2013, January). An ontology of scientific uncertainty: methodological lessons from analyzing expressions of uncertainty in food risk assessment. In *SAESA Amsterdam Conference.*

Bongelli, R., Riccioni, I., Burro, R., & Zuczkowski, A. (2019). Writers' uncertainty in scientific and popular biomedical articles. A comparative analysis of the British Medical Journal and Discover Magazine. *PLoS One* , *14*(9), e0221933.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In *Psychology of learning and motivation* (Vol. 32, pp. 275-318). Academic Press.

Candlin, C. N., & Hyland, K. (Eds.). (2014). *Writing: Texts, processes and practices*. Routledge.

Candlin, F. (2000). Practice-based doctorates and questions of academic legitimacy. *Journal of Art & Design Education*, *19*(1), 96-101.

Chen, C., Song, M., & Heo, G. E. (2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics*, *12*(1), 158-180.

Cordner, A., Brown, P.: Moments of uncertainty: Ethical considerations and emerging contaminants. *Sociological Forum*, 28(3), 469–494 (2013) https://doi. org/10.1111/socf.12034

Courtney, H. (2001). *20/20 Foresight: Crafting strategy in an uncertain world*. Harvard Business Press.

Dosi, G., & Egidi, M. (1991). Substantive and procedural uncertainty: an exploration of economic behaviours in changing environments. *Journal of evolutionary economics*, *1*, 145-168.

Salager-Meyer, F. (2017). I Think That Perhaps You Should: A Study of Hedges in Written Scientific.

Fijnvandraat, M. L. (2009). Shedding light on the black hole: The roll-out of broadband access networks by private operators. (Doctoral dissertation, Delft University of Technology). Next Generation Infrastructure Foundation. ISBN 978-90-79878-03-1

Fox, Craig R. and Gülden Ülkümen (2011), Distinguishing Two Dimensions of Uncertainty. In *Essays in Judgment and Decision Making*, Brun, W., Kirkebøen, G. and Montgomery, H., eds. Oslo: Universitetsforlaget. http://dx.doi.org/10.2139/ssrn.3695311

Friedman, M., & Kandel, A. (1999). *Introduction to pattern recognition: statistical, structural, neural, and fuzzy logic approaches* (Vol. 32). World scientific.

Funtowicz, S. O., & Ravetz, J. R. (1990). *Uncertainty and quality in science for policy* (Vol. 15). Springer Science & Business Media.

Guillaume, J. H., Helgeson, C., Elsawah, S., Jakeman, A. J., & Kummu, M. (2017). Toward best practice framing of uncertainty in scientific publications: A review of W ater R esources R esearch abstracts. *Water Resources Research*, *53*(8), 6744-6762. https://doi.org/10.1002/2017WR020609.

Helton, J. C. (1994). Treatment of uncertainty in performance assessments for complex systems. *Risk analysis*, *14* (4), 483-511.

Hyland, K. (1998). Hedging in scientific research articles. *Hedging in Scientific Research Articles*, 1-317. Amsterdam: John Benjamins. http://dx.doi.org/10.1075/pbns.54

Hyland, K. (2000). Developments in English for Specific Purposes: A multi-disciplinary approach. *English for specific purposes*, *19*(3), 297-300.

Hyland, K. (1996). Talking to the academy: Forms of hedging in science research articles. *Written communication* , *13*(2), 251-281.

Jauch, L. R., & Kraft, K. L. (1986). Strategic management of uncertainty. *Academy of management review*, *11*(4), 777-790.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition*, *11*(2), 143-157.

Light, M., Qiu, X. Y., & Srinivasan, P. (2004). The language of bioscience: Facts, speculations, and statements in between. In *HLT-NAACL 2004 workshop: linking biological literature, ontologies and databases* (pp. 17-24). https://aclanthology.org/W04-3103

Medlock, B., & Briscoe, T. (2007, June). Weakly supervised learning for hedge classification in scientific literature. In *ACL* (Vol. 2007, pp. 992-999).

Meijer, I. S., Hekkert, M. P., Faber, J., & Smits, R. E. (2006). Perceived uncertainties regarding socio-technological transformations: towards a framework. *International Journal of Foresight and Innovation Policy*, *2*(2), 214-240.

Morgan, M. G., Henrion, M., & Small, M. (1992). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge university press.

Ningrum, P.K., Atanassova, I. (2023). Dataset for Multidisciplinary Uncertainty Mining - Ver1. https://doi.org/10.5281/zenodo.8024787

Ningrum, P. K., Mayr, P., & Atanassova, I. (2023). UnScientify: Detecting Scientific Uncertainty in Scholarly Full Text. *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/2307.14236

Pinto, M. D. G., Osorio, P., & Martins, F. (2014). A theoretical contribution to tackling certainty and uncertainty in scientific writing: four research articles from the journal Brain in focus. *Communicating Certainty and*

Uncertainty in Medical, Supportive and Scientific Contexts. Amsterdam: John Benjamins Publishing Company, 291-308.

Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., & Pavel, M. (2005). A typology for visualizing uncertainty. *Proceedings of SPIE - The International Society for Optical Engineering*, *5669*, 146-157. Article 16. https://doi.org/10.1117/12.587254

Rammel, C., & van den Bergh, J. C. (2003). Evolutionary policies for sustainable development: adaptive flexibility and risk minimising. *Ecological economics*, *47*(2-3), 121-133.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process–a framework and guidance. *Environmental modelling & software*, 22(11), 1543-1556. https://doi.org/10.1016/j.envsoft. 2007.02.004

Rey, F. C., Bertin, M., & Atanassova, I. (2018, June). Une étude de l'incertitude dans les textes scientifiques: vers la construction d'une ontologie. In *TOTh 2018 Terminology & Ontology: Theories and applications* (pp. 229-242). Presses universitaires Savoie Mont Blanc.

Rubin, V. L., Liddy, E. D., & Kando, N. (2006). Certainty identification in texts: Categorization model and manual tagging results. *Computing attitude and affect in text: Theory and applications*, 61-76.

Stocking, S. H., & Holstein, L. W. (1993). Constructing and reconstructing scientific ignorance: Ignorance claims in science and journalism. *Knowledge*, *15*(2), 186-210.

Swales, J. M. (2014). 1990. Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press, selected 45--47, 52--60. In *The Discourse Studies Reader* (pp. 306–316). Amsterdam: John Benjamins Publishing Company.

Van Asselt, M. (2000). *Perspectives on uncertainty and risk: the PRIMA approach to decision support*. Springer Science & Business Media.

Van Asselt, M., & Rotmans, J. (2000). Uncertainty in integrated assessment, A bridge over troubled water. *ICIS (International Centre for Integrative Studies) Maastricht University*, *60*.

van der Bles, A. M., van der Linden, S., Freeman, A., & Spiegelhalter, D. (2018). 18 The effects of communicating uncertainty about facts and numbers. *Evidence-Based Medicine*, 23 (Suppl 1), pp. A9–A10.

van der Sluijs, J. P. (1997). Anchoring amid uncertainty. *On the management of uncertainties in risk assessment of anthropogenic climate change. CIF-Gegevens Koninklijke Bibliotheek, Den Haag*.

van Witteloostuijn, A. (1986). *Choice-theory versus vs. behaviourism: a paradox*. Groningen: University of Groningen, pp.1–16.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation, and their scopes. *BMC bioinformatics*, *9*(11), 1-9. https://doi.org/10.1186/1471-2105-9-S11-S9

Walker, W. E., Harremoës, P., Rotmans, J., Van Der Sluijs, J. P., Van Asselt, M. B., Janssen, P., & Krayer von Krauss, M. P. (2003). Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated assessment*, 4(1), 5-17. https://doi.org/10.1076/iaij.4.1.5.16466

Yao, M., Wei, Y., & Wang, H. (2023). Promoting research by reducing uncertainty in academic writing: A large-scale diachronic case study on hedging in Science research articles across 25 years. *Scientometrics*, *128*(8), 4541–4558. https://doi.org/10.1007/s11192-023-04759-6