Scientific Uncertainty: An Annotation Framework and Corpus Study in Different Disciplines

Panggih Kusuma Ningrum, Iana Atanassova

Scientific Uncertainty: An Annotation Framework and Corpus Study in Different Disciplines

Panggih Kusuma Ningrum[1] and Iana Atanassova[2]

[1] *panggih_kusuma.ningrum@univ-fcomte.fr*
Université de Franche-Comté, CRIT, F-25000 Besançon, France

[2] *iana.atanassova@univ-fcomte.fr*
Université de Franche-Comté, CRIT, F-25000 Besançon, France
Institut Universitaire de France (IUF), France

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 categories sentences containing uncertainty along five dimensions. We study a corpus of articles from different disciplines and conduct experiments on two different samples of sentences: one sample extracted by uncertainty cue mapping and another sample obtained from manual annotation of randomly selected articles. The two samples are manually annotated using our annotation framework. The results show the distribution of uncertainty types across journals and categories. The samples of annotated sentences can also be used to automate some aspects of the annotation process.

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 & 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 provide evidence on the types of scientific uncertainty mobilised in different disciplines and their positions as part of scientific discourse. To this end, we construct a dataset of articles and observe the types of uncertainty expressed in two different samples of sentences: a sample extracted by uncertainty cue mapping, and another sample obtained by manual annotation of randomly selected articles. A secondary goal of this research is to construct an annotated dataset that can be used to implement automated tools[1].

The following sections of this paper are organised as follows: The first section provides a comprehensive review of relevant research studies investigating the classification and identification of scientific uncertainty. The next section outlines the methodology employed, including the selection of the dataset and an introduction to the annotation framework. It also provides an overview of the research pipeline for the two experiments - Uncertainty Cue

[1] The dataset will be published in Open Access on the Zenodo platform if the article is accepted for publication.

Mapping and Manual Uncertainty Expression Search. Following this, the results section provides a detailed account of the frequencies and distributions of uncertainty expressions across different categories and journals for the two experiments. Finally, a discussion of these results is presented.

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 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. Journal articles are becoming a new alternative to the traditional method of communicating new findings and ideas, which scientists and academics used to use via letters. Most importantly, journal articles now 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 inherent in the vast and rapidly growing volume of journal articles remains a barrier (Chen, Song, & Heo, 2018). The fundamental problem is that working with unstructured textual data in the scientific literature is challenging. Most previous studies have focused on detecting and identifying a specific set of uncertainty cues and markers in scientific articles by using a specific part of the text, such as the abstract (Vincze et al., 2008; Guillaume et al., 2017) or the full text (Hyland, 1996; Medlock & Briscoe, 2007). These studies have contributed to the expansion of the uncertainty vocabulary and lexicon, but the implementation of the technique is often misleading due to the high complexity of natural language.

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 (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 et al. (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 & Ravetz, 1990; Morgan & Henrion, 1992; Van Asselt, 2000; Van der Sluis, 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 (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 our study includes scientific uncertainty, which is expressed in journal articles, whereas Walker included external factors in the framework, such as stack holders 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.

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 t. This framework is intended to be interdisciplinary. An uncertainty categorisation model with five dimensions: Reference, Nature, Context, Timeline and Expression, is proposed. Figure 1 shows these five dimensions and how each dimension is subdivided into categories. A detailed description of the dimensions is provided in the following sections.

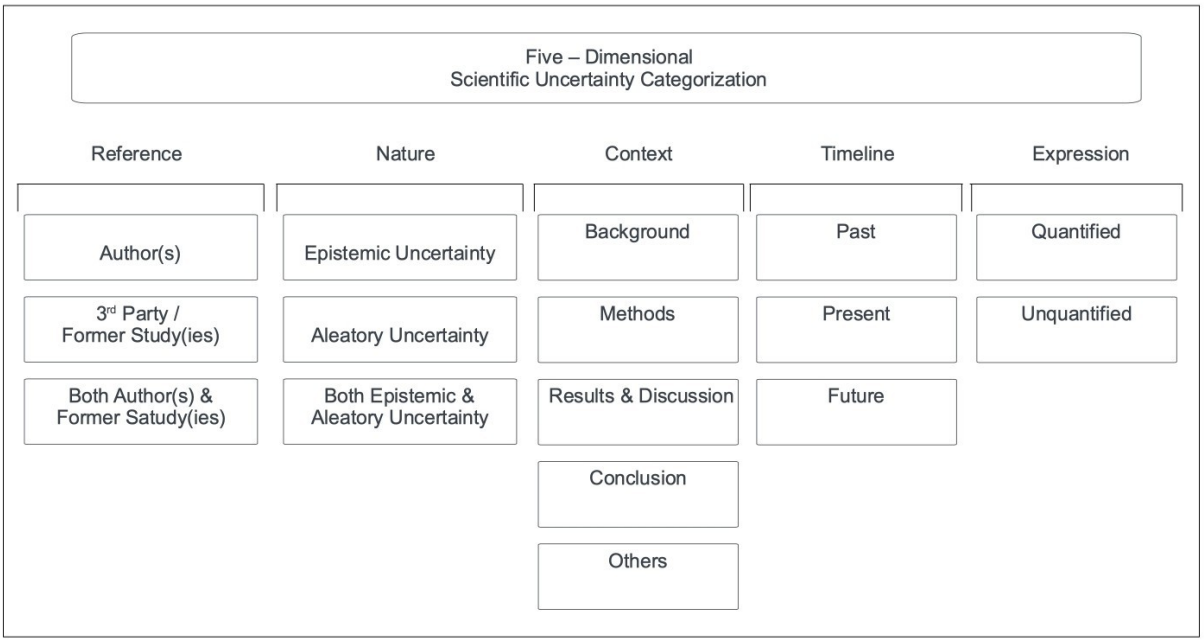

Figure 1: Framework for Scientific Uncertainty Categorization

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 is intended to accommodate complex sentences that may refer to both the author(s) and the previous study(s).

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 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 & Egidi 1991; Helton, 1994; Jauch & Kraft, 1986; Kahneman & Tversky, 1982; Rammel & Bergh, 2003; Van Asselt & Rotmans, 2000; van der Sluijs, 1997; Walker et al., 2003).
- 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 & Egidi 1991; Helton, 1994; Jauch & Kraft, 1986; Kahneman & Tversky, 1982; Rammel and Bergh, 2003; Van Asselt & Rotmans, 2000; van der Sluijs, 1997; Witteloostuijn, 1986; Walker et al., 2003).

Additionally, the last group in this category is for complex sentences that may consist of the combination of Epistemic and Aleatory 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).

Expression.

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

- 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	Reference	Nature	Context	Timeline	Expression
In addition, results of in-vitro experiments indicate that statins might inhibit bone resorption by interfering with osteoclast function in a similar way as bisphosphonates [16].	BMC Med	Former studies	Aleatory	Background	Past	Unquantified
These results suggest that a significant portion of TLR9 may avoid relocalization in Unc93b1-Ist2 cells and signal from internal compartments.	Nature	Author(s)	Epistemic	Background	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	Result & discussion	Future	Unquantified

Dataset

Data Selection

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.

Corpora Construction

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. For this reason, three corpora were created, consisting of journals from (1) medicine, (2) biochemistry, genetics, and molecular biology, and (3) multidisciplinary journals. The main purpose of this is to observe the differences in the characteristics of uncertainty expressions in each discipline.

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 of its subject area. Next, the top two journals were selected for the Medicine and Biochemistry corpus, while for the Multidisciplinary corpus, PLoS One and Nature were selected as these two journals met the data selection criteria. In addition, they are also indexed in Scopus and Web of Science in the first quartile (Q1) for multidisciplinary field and have a large repository. Table 2 describes the list of journals and the distribution of data in the corpora.

Table 2. Corpora Description

Discipline	Journal	Total Articles	Total Sentences
Medicine	BMC Med	535	93 700
	Cell Mol Gastroenterol Hepatol	586	176 597
Biochemistry, Genetics & Molecular Biology	Nucleic Acids Res	1 871	312 492
	Cell Rep Med	263	89 652
Multidisciplinary	Nature	831	108 153
	PLoS One	322	54 336
	Total	*4 408*	*834 930*

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 hints 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.

Research Pipeline

Five 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, and (5) Annotation. Three data are used as the inputs such as Lists of uncertainty cues and markers from (Hyland 1996; Chen, Song, and Eun Heo 2017; Bongelli et al. 2019), scientific articles that are stored on a MySQL database, and the Five-Dimensional Scientific Uncertainty Categorization. Figure 2 describes the stages involved in this study.

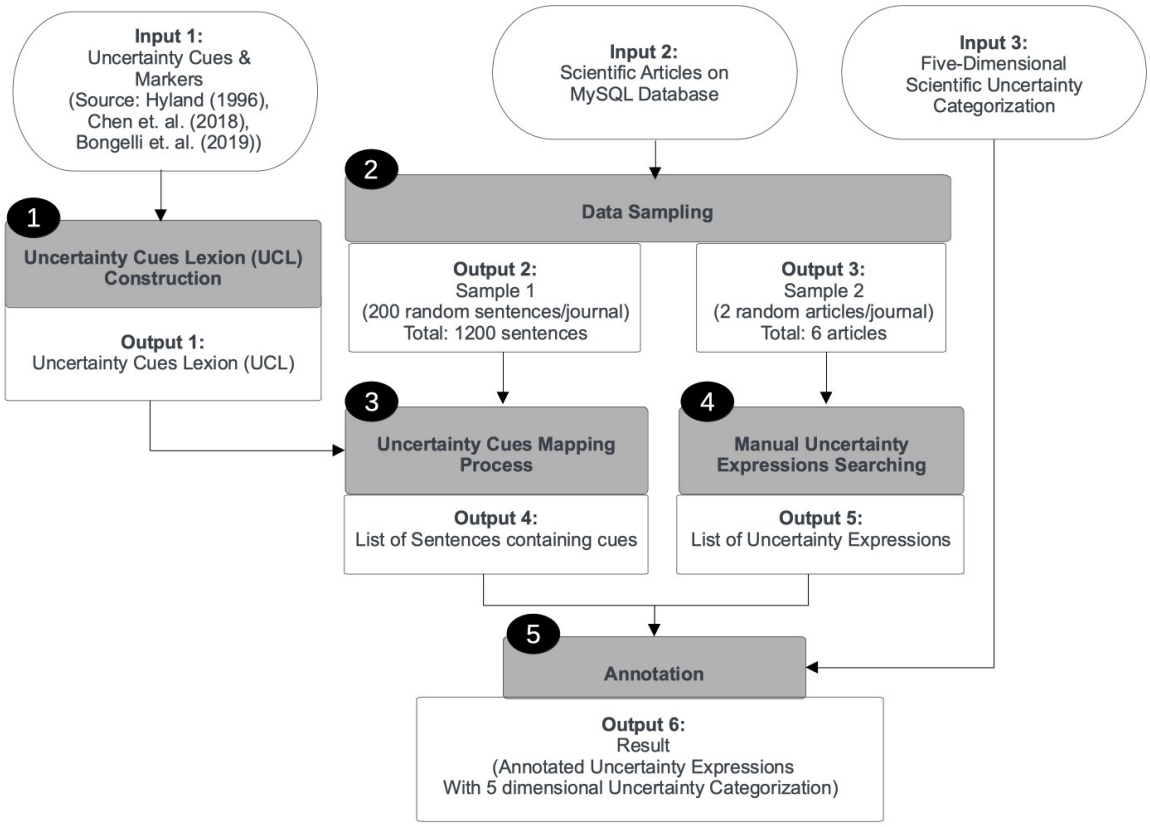

Figure 2. Research Path Diagram

Uncertainty Cues Lexicon (UCL) construction

As illustrated in Figure 2, a list of cues and markers expressing uncertainty from Hyland (1996), Chen, Song, and Eun Heo (2017), and Bongelli et al. (2019) were adopted and compiled to create the Uncertainty Cues Lexicon (UCL). The list of cues is shown in Table 3.

Table 3. 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

Data Sampling

Two methods of data sampling were employed in this study.
Firstly, 200 sentences were randomly selected from the MySQL database for each target journal. This first sample (Sample 1) of 1200 sentences was used as input for the uncertainty cue mapping process.
Secondly, two articles were randomly selected from each targeted journal. This second sample (Sample 2) of 12 articles was used for the manual search for uncertainty terms.

Uncertainty Cues Mapping Process

Sample 1 and the UCL were used as inputs for this step. Regular Expressions (RegEx) were used to perform the uncertainty cue mapping process. In practice, each cue in the UCL was mapped to Sample 1 and then all sentences containing cues were listed.

Manual Uncertainty Expressions Searching Process

In this step, a manual search was carried out using Sample 2 as data. Every sentence in each article was screened and the sentences expressing uncertainty were marked. The marked sentences were then compiled into a list and prepared for the annotation process.

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 "No Uncertainty" or containing “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.

Results

In this section, we present the results of scientific uncertainty identification and categorisation from two experimental settings, namely Uncertainty Cue Mapping and the Manual Uncertainty Expression Searching process.

Uncertainty Cue Mapping in Scientific Articles

In the overall sample of 1200 sentences, 258 sentences (21.50%) were detected as containing uncertainty cues. Among them, 107 sentences (8.92%) were annotated as expressing uncertainty. Table 4 describes the results of the uncertainty cue mapping process in more details. Among the journals, BMC Medicine (32) contributes to the highest number of the 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).

Table 4. Results of cue mapping on the total sample of 1200 sentences

Discipline	Journal	Articles with cue(s)	Sentences with cue(s)	Cue occurences*	Perc. of sentences with cue by total samples	Ucertainty occurrences (sentences)	Perc. of uncertainty occurrences by total samples	Perc. of uncertainty occurrances by total cues
Medicine	BMC Med	49	58	84	29,00%	32	16,00%	38,10%
	Cell Mol Gastroenterol 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%
Multidisciplinary	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

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)

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.

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). Figure 4 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.

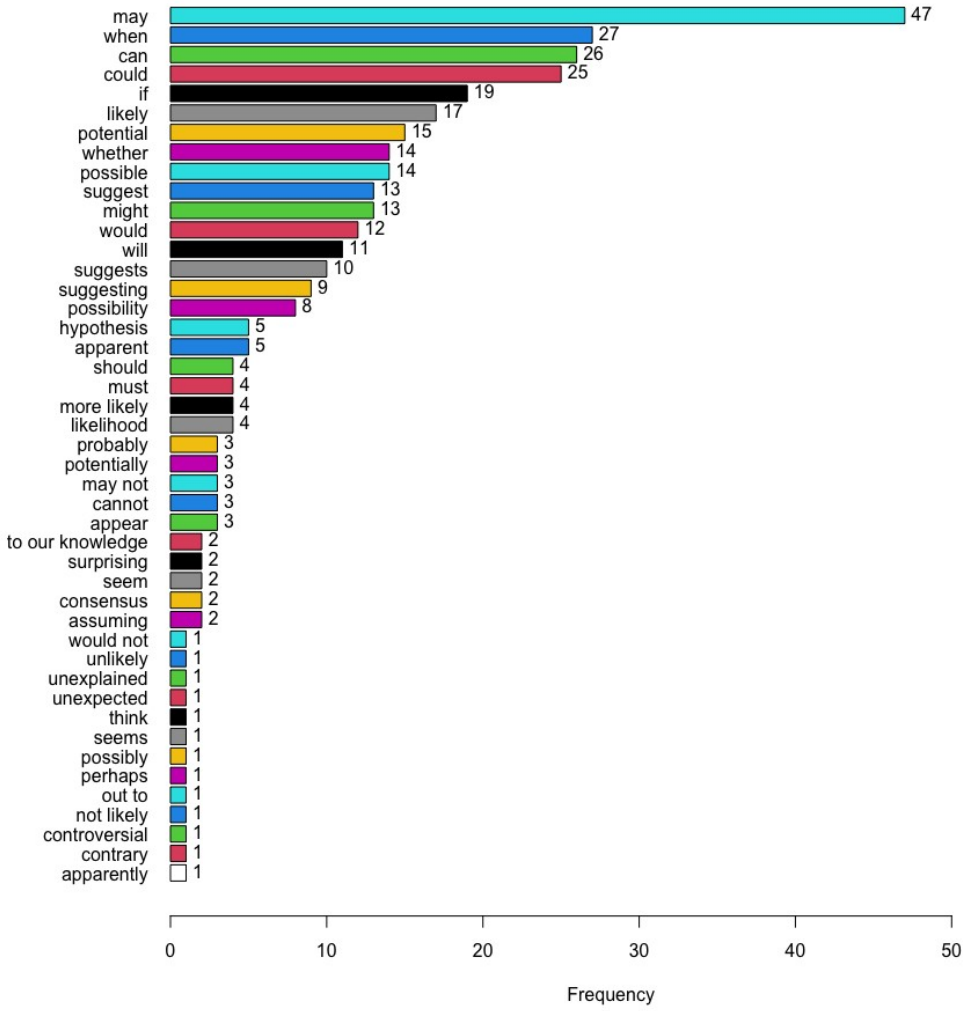

Figure 3. Uncertainty cues occurrences

Table 5 presents the distribution of sentences expressing uncertainty in the different categories. In each dimension, we observe some very important differences between the categories. In terms of Reference, the vast majority of uncertainties (87.0%) were annotated as "Author(s)", while only 10.5% were annotated as "Previous studies" and 2.5% were annotated as both. Most sentences express Epistemic uncertainty (77.2%), while only 22.8% are Aleatory. The Context dimension indicates the section in which the sentence appears in the article. More than half of the uncertainties are expressed in the Results and Discussion section (57.04%), while the other sections contribute to a lesser extent. The second section with the most uncertainties is Others (21.6%) following by the Background section (17.9%). In terms of Timeline, the Past and Future categories are rare (less than 20% in total), while the Present accounts for 81.5%. Finally, the large majority of uncertainties are Unquantified, with only 0.6% Quantified.

Table 5. Uncertainty distribution by categories (Cue Mapping Results)

Uncertainty Category		Proportion in each Category (%)
Reference	Author(s)	87.0%
	Former Study(s)	10.5%
	Both	2.5%
Nature	Epistemic	77.2%
	Aleatory	22.8%
	Both	0,00%
Context	Background	17.9%
	Method	2.5%
	Results & Discussion	57.4%
	Conclusion	0.6%
	Others	21.6%
Timeline	Past	15.4%
	Present	81.5%
	Future	3.1%
Expression	Quantifiable	0.6%
	Unquantifiable	99.4%

Manual Uncertainty Expression Searching in Scientific Articles

In the sample of 12 articles, a total of 95 sentences were annotated with occurrences of uncertainty. Table 6 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.

Table 6. Results of Manual Searching

Discipline	Journal	Uncertainty Occurrances in sentence
Medicine	BMC Med	36
	Cell Mol Gastroenterol Hepatol	5
Biochemistry, Genetics & Molecular Biology	Nucleic Acids Res	13
	Cell Rep Med	19
Multidisciplinary	Nature	14
	PLoS One	8

Table 7 displays the distribution of sentences expressing uncertainty across the different categories for this second sample. As before, it can be seen that the majority (88.4%) of the sentences are annotated as Author in the Reference category. In terms of Nature, 81.1% of the sentences are annotated as Epistemic, while the rest are Aleatory or Both. About 46% of the uncertainties are found in the Result and Discussion sections. In terms of Timeline, again the majority is annotated as Present (76.8%), and a relatively small number of sentences are annotated as Past and Future. Finally, in this sample all sentences are annotated as

Unquantified and no occurrences of Quantified were found. Overall, this distribution is quite similar to the one observed for Sample 1 (see Table 5).

Table 7: Uncertainty distribution by categories (Manual Searching Results)

Uncertainty Category		Proportion in each Category (%)
Reference	Author(s)	88.4%
	Former Study(ies)	8.4%
	Both	3.2%
Nature	Epistemic	81.1%
	Aleatory	16.8%
	Both	2.1%
Context	Background	18.9%
	Method	3.2%
	Results & Discussion	46.3%
	Conclusion	6.3%
	Others	25.3%
Timeline	Past	14.7%
	Present	76.8
	Future	8.4%
Expression	Quantifiable	0%
	Unquantifiable	100%

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 respects. First, the size of the two samples we examined is relatively small (1200 sentences and 12 articles), which could lead to over- or under-representation of some categories. We plan to carry out studies with larger samples. However, the human effort required for this kind of annotation is important, as each sentence must be carefully examined and annotated according to five dimensions. Secondly, the disciplines and journals selected are small, only 2 disciplines and 2 multidisciplinary journals. 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 the two multidisciplinary journals do not contain enough sentences to observe this.

The sampling methods were chosen in a way that existing resources (cue lists from previous studies) are exploited 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.

Conclusion and future work

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 sentences from different disciplines that were manually annotated. 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.

Acknowledgements

This work is supported by the ANR InSciM Project (2021-2024), funded by French ANR, grand number ANR-21-CE38-0003-01.

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.

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. (2013, September). Moments of uncertainty: Ethical considerations and emerging contaminants. In *Sociological Forum* (Vol. 28, No. 3, pp. 469-494).

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.

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, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. *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.
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).
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.
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.
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.
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.
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.

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.