Studying Uncertainty in Science: a distributional analysis through the IMRaD structure

Iana Atanassova, François-C. Rey, Marc Bertin

Studying Uncertainty in Science: a distributional analysis through the IMRaD structure

Iana Atanassova[1], **François-Claude Rey**[1] **and Marc Bertin**[2]
[1] CRIT - Centre Tesnière, Université de Bourgogne Franche-Comté
30 rue Mégevand 25000 Besançon, France
iana.atanassova@univ-fcomte.fr, francois_claude.rey@edu.univ-fcomte.fr
[2] ELICO laboratory, University Claude-Bernard Lyon-1
43 Bd du 11 novembre 1918, 69622 Villeurbanne cedex, France
marc.bertin@univ-lyon1.fr

Abstract

Uncertainty in science is an integral part of the research process and an important element of discovery. We propose to study the expression of uncertainty in research papers in two domains which are Biomedicine and Physics. We focus on two existing sets of cue words and construct a new set of strong indicators of uncertainty that are composed of linguistic expressions. After processing two datasets that consist of the full text of papers published by PubMed in Open Access, we examine the occurrences of the sets of cue words and strong indicators. As a result, we obtain the distributions of cues of uncertainty along the text progression and their positions with respect to section types in the rethorical structure of the papers. The results show that cues and strong indicators of uncertainty are most frequent in the Discussion section of articles and, in general, towards the end of the text progression. Significant differences can be observed between the two datasets.

Keywords: scientific papers, uncertainty, cue words, strong indicators, hedging, IMRaD, distributional analysis

1. Introduction

1.1. Uncertainty in Science as a Subject of Study

The study of uncertainty in science represents a fundamental challenge for the understanding of science, as a study of its social structures and also of what regards its relation with other domains or institutions. This problem, uncertainty and the effects of uncertainty, as an object of study, is a relatively old one. (Friedman et al., 1999) explores scientific uncertainty and media coverage of science in major public issues such as AIDS, biotechnology, dioxin, global warming, and nature. Uncertainty concerns all scientific fields, their funding, their perception in the public opinion, their relation with the legal system and the discussions revolving around environment and climate change, etc. For example, there is a major concern about using paradigm development as a measure of uncertainty in science, as shown by (Pfeffer et al., 1976). In physics and chemistry, uncertainty is both a practical and a theoretical notion that is taken into consideration in the construction of new knowledge. Some studies focus on the relationship between data quality in analytical chemistry and uncertainty, e.g. (Committee, 1995) shows that a measurement cannot be properly interpreted without the knowledge of its uncertainty.

The perception of uncertainty is an important issue of all scientific activities, and beyond, in the fields of health, climate and environment. Indeed, in health risk assessments, it is an important element in terms of public health and a subject of study in itself as shown by (Johnson and Slovic, 1995), and it could have important implications in the communication of risks. Some empirical studies, e.g. (Flemming et al., 2015), show that certain aspects of uncertainty influence the understanding of the uncertainty of scientific information and the perception of credibility. Concerning the debate around global warming, uncertainty is an object of study related to the validity of certain numerical simulations, and is at the heart of many discussions on the simulations of future climate changes. Nevertheless, this debate affects the perceived authority of science and challenges the authority of climate science, mainly in the context of policy development (Shackley and Wynne, 1996). Some literature reviews illustrate the problems and inconsistencies in conceptualizing and measuring construction in the environmental domain, designating three types of uncertainty and describing their implications, e.g., (Milliken, 1987).

The uncertainty in policy knowledge is a central element of many policies based on scientific knowledge and the precautionary principle (Wynne, 1992). Understanding scientific uncertainty is essential to inform in the debate between the advocates of the "precautionary principle" and those of the "scientific regulation". This debate is based on the notion of the standard of proof to be applied to scientific evidence that a given action presents a danger of seriousness and irreversibility, e.g. (Weiss, 2003).

Uncertainty is a part of science because it is related to the goals of science when science aims at knowing the unknown. It is attached to the objects of science because they depend on precising and sharing their definitions, and even the concepts of uncertainty are defined in each scientific field according to the needs of that field : e.g. the "measurement error" in physics (Joint Committee for Guides in Metrology, 2012), the "proof levels" in biomedicine (HAS - Haute Autorité de la Santé, 2013), and the "incompleteness" or "unpredictability" in humanities and social sciences (Fusco et al., 2015). Furthermore, uncertainty is a part of the methods, tools and scientific results when science removes ambiguities. In fact, the results of science depend generally on the limits of precision of measurement methods and tools. For example in Physics, (Joint Committee for Guides in Metrology, 2008) points out that, depending on what should be the use of the measure of the result,

"when reporting the result of a measurement and its uncertainty, it is preferable to err on the side of providing too much information rather than too little", and pointing out that one should, among others things, "list all uncertainty components and document fully how they were evaluated". Uncertainty is a part of science through the interpretations and communication of its results. For example, it is connected with the precautionary principle, e.g. in economics, when public authorities cope with environmental and health risks, they might want to adjust decisions between, on the one hand, environmental protection or damage and, on the other hand, economic development or slowdown. They could look for economic tools in order to find a strategy which will maximise expectation and minimise risk. However, experts interpretations or recomendations remain sometimes controversial due to different schools of thought or choices of concepts.

Uncertainty is an international topical matter, e.g. with the current problem of climate change, because the scope of the precautionary principle "covers those specific circumstances where scientific evidence is insufficient, inconclusive or uncertain and there are indications through preliminary objective scientific evaluation that there are reasonable grounds for concern (...)."[1]. Climate models, i.e. "computer simulations based on physical laws of atmospheric conditions" (Caers, 2011, p. 50-52), explore possibilities limited e.g. by the computer's power to calculate equations. (World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), 2005, p. 26) also reports that "risk assessment regarding, for instance, anthropogenic climate change [...], involves uncertainties of many sorts, not all of which can be resolved.", but "high quality science does not require low uncertainty.".

Following these points of view, and (Fusco et al., 2015), it appears that uncertainty is one of the common components in the results of scientific research, whose dimensions can be multiple in the various disciplinary fields that appropriate it. This component must be reusable with the scientific results that it accompanies and of which it is an essential integral part. If in Humanities and Social Sciences (HSS) statements based on objectivity or subjectivity are present, the concept of uncertainty is extended: inaccuracy, indeterminacy, incompleteness, ambiguity and unpredictability. Uncertainty in HSS is related to the complexity of social and human study objects, to the influence of the context and methods, to the diverse perspectives and paradigms, to the controversies and variety of viewpoints and interpretations, and to the mode of communication of the results. It is then a question of "*making science with uncertainty*".

From a definitional point of view; in philosophy *subjectivity* is opposed to *objectivity* and considered as quality (unconscious or inner) of what belongs only to the thinking subject; whereas in linguistics it is related to the presence of the speaking subject in his speech[2]. Studied respectively by logic with the square of Aristotle's modal logics, by philosophy with Kant, and finally by linguistics, the concepts related to subjectivity are complementary and sometimes non-tunable, even within the same disciplinary field. In linguistics, (Bally, 1965), doing "the logical analysis of the forms of enunciation", considers that the explicit sentence includes two complementary parts which are: the "dictum" as "the correlative of the process which constitutes the representation, and the "modus" as "the expression of the modality, correlative to the expression of the thinking subject" consisting of a modal verb and a modal subject. Concepts and relationships around the concept of modality are not consensual as evidenced by many works.

In this paper we are interested in the expression of uncertainty in scientific papers. For this purpose, we adopt the following definition:

Definition: *A sentence in a scientific article expresses uncertainty if there exists explicit evidence that the author modifies the epistemic value of the proposition.*

1.2. Research Problem

Recognizing, collecting and presenting subjectivity and uncertainty is of considerable importance in scientific activities, since it allows one to know, with greater acuity and perspective, what are the driving forces, perceptions and trends of development of research in science course. This is useful for supplementing the information stated in a strict framework, but also for understanding the motivations of researchers and the creative aspects of research.

A tool that allows to identify and classify these phenomena in texts would have many applications for the researchers, in order to: produce and to supplement states of the art, analyze a disciplinary field, analyze the temporal scale of the researches and the decidability of the technical and economic prospective, and automate the reconstruction of the genesis of scientific concepts over a period by a diachronic approach.

The identification of textual segments expressing uncertainty has been the goal of several studies. The identification of speculative sentences in texts by approaches in machine learning was addressed by (Moncecchi et al., 2012) which underlines the specificity of the problematic of subjectivity. The identification of uncertainty through hedging was the main objective of the CoNLL-2010 shared task[3] (Farkas et al., 2010). A recent study by (Chen et al., 2018) proposes to identify introductory expressions of uncertainty by expanding a restricted set of expressions. However, these last works express a binary view of uncertainty and do not address the different levels and dimensions of uncertainty in order to account for the complexity of this notion. (Bernhard and Ligozat, 2011) use certainty gradations based on categories to capture assertions associated with information about medical problems in clinical reports. A related work on the detection of different types of hypotheses in Biomedical papers has been conducted by (Desclés et al., 2014; Desclés et al., 2009).

A related topic is the detection of hedging, which is not only limited to cases of uncertainty, but also for example as "a service to the reader", or the so called reader-oriented

[1] "Communication from the Commission on the precautionary principle COM/2000/0001 final", URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52000DC0001

[2] Source: National Center for Textual and Lexical Resources, France - CNRTL

[3] http://rgai.inf.u-szeged.hu/conll2010st/tasks.html

hedges (Hyland, 1996b). In our study we focus particularly on uncertainty, hedging being a much broader phenomenon.

In this paper we focus on another aspect of the expression of uncertainty: its position in the rethorical structure of articles. To our knowledge, no other works have studied this phenomenon on a large-scale dataset.

Our goal is to propose a linguistic model of the expression of uncertainty in scientific articles. We study the definitions of the notions of subjectivity and uncertainty as well as their expression in the different scientific disciplines in order to lead to the creation of an ontology of uncertainty in science. The purpose of this work is the creation of a tool for the identification and automatic extraction of segments expressing uncertainty in scientific articles. This research is part of a series of works around the analysis of the content of scientific articles with the aim of automating the state of the art of scientific domains.

We propose to study the expression of uncertainty by examining the occurrences of cue words in a large corpus of scientific papers in several different journals. Firstly, we use two core sets of cue words proposed by two previous studies: H-set (Hyland, 1996a) and C-set (Chen et al., 2018), and we examine the sentences containing these cue words. Secondly, to be able to identify sentences expressing uncertainty with a better precision we propose a new set of strong indicators of uncertainty that we call A-set. Concerning these three sets of cue words and expressions, we focus on two major phenomena that we study and describe in detail:

1. their distribution in the text progression of papers;
2. their positions in the structure of papers wich respect of section types.

The rest of this paper is organised as follows. The next section describes the datasets, the sets of cues and indicators that were considered and the overall methodology for the distributional analysis. Section 3 presents the results of the experimentation on the datasets and the obtained distributions. Finally, section 4 proposes a discussion and concludes the article.

2. Methods

2.1. Datasets

We have processed two datasets of papers published in journals in the fields of Biomedicine and Physics. The Biomed dataset contains 9 463 papers from 7 different journals, and the Physics dataset contains 488 papers from 2 journals. These datasets are part of the PubMed Open Access Subset[4]. We have downloaded the entire collection of research articles published in Open Acces for each of the journals. The articles are in the XML JATS format, where the body consists of sections and paragraphs that are identified as distinct XML elements. We have carried out the sentence segmentation of all the text. Table 1 shows the journals that were included in the datasets, together with the number of papers and the average size of the papers in each journal in terms of number of sections and number of sentences.

[4]https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

The two datasets are different in size and include papers of various lengths (from 124 to 284 sentences on average). Our objective in this study is to observe the expression of uncertainty and its position in the text of papers. Concerning the Biomed dataset, we are interested in the variations that we can observe between the different journals, as well as in any shared properties between them. The Physics dataset is much smaller and will allow us to make a first comparison between uncertainty as it is expressed in the Biomedical field and in Physics. However, we note that the size of the Physics dataset does not allow to draw any important conclusions for this field.

2.2. Sets of uncertainty cues

From a linguistic point of view, the expression of uncertainty has been studied through the phenomenon of hedging (Lakoff, 1973). (Hyland, 1996a; Hyland, 1998) addresses the problem of hedging in scientific articles, and examines the frequencies of modal verbs (*would (not), may (not), could*) and epistemic lexical verbs (*indicate, suggest, appear, ...*) in different corpora. (Chen et al., 2018) study the set of cues proposed by Hyland and in addition to it compile manually a new set of 27 cue words that include mainly nouns and adjectives such as *unclear, controversial, inconclusive*. In our experiment, we have considered these two sets of cues that we call H-set and C-set. Table 2 gives the complete lists of cues in these two sets.

We note however that the simple presence of such cues in a sentence does not necessarily imply that the sentence expresses uncertainty. Natural languages make use of complex lingsuitic operators and the meaning of an expression or a sentence is different from the sum of the meanings of the words that compose it. Table 3 shows examples of sentences that contain cues from the C-set but do not express uncertainty.

2.3. Construction of the A-set of strong indicators of uncertainty

The examples on table 3 indicate that the task of the identification of sentences that express uncertainty in scientific texts cannot be tackled by the simple identification of cue words. This is due to the fact that most of these cues are highly amibiguous and can occur in a variety of contexts that do not necessarily express uncertainty, leading to noise in the classification.

In fact, the identification of one-word cues allows to identify sets of sentences that might express uncertainty, but the presence of such a cue in itself is not sufficient to classify the sentence as an expression of uncertainty. At the same time, the idea of uncertainty is most often conveyed by expressions which are composed of several words that can be non contiguous in the sentence. For this reason, we need to consider sets of more complex expressions, in order to be able to identify sentences containing uncertainty with a higher precision, and to limit the noise. Table 4 shows examples of sentences that express uncertainty and where the meaning of uncertainty is carried by an expression larger than one word. Some of these sentences contain also uncertainty cues from the H-set and the C-set.

To tackle such problems by using linguistic resources, vari-

Journal Name	Nb of papers	Avg nb of sections	Avg nb of sentences
Biomed dataset			
International Journal of Genomics	246	4,92	145,98
Physiological Reports	1 667	5,33	165,09
Stem Cell Research & Therapy	881	6,00	185,22
Cancer Science	865	4,10	123,94
Cell Death & Disease	2 742	3,08	165,64
ZooKeys	1 710	4,59	269,68
Microbial Cell Factories	1 352	5,92	187,94
Total Biomed	*9 463*	*4,57*	*185,03*
Physics dataset			
Journal of Nanoparticle Research	118	4,77	190,27
The European Physical Journal. C, Particles and Fields	370	7,18	283,63
Total Physics	*488*	*6,59*	*261,06*

Table 1: Datasets description.

H-set (Hyland, 1996a)	C-set (Chen et al., 2018)	
would (not)	unclear	suspect
may (not)	controversial	ambiguity
could	inconclusive	unexpected
might (not)	consensus	contrary
should	inconsistent	paradoxical
cannot	confusing	unusual
will (not)	uncertain	flaw
must	uncertainty	dispute
shall	unknwown	impossible
ought to	ambiguous	misleading
	incomplete	unexplained
	contradictory	contentious
	paradox	incompatible
	surprising	

Table 2: Lists of hedging cues: H-set and C-set.

ous knowledge-based methods have been proposed, e.g. the microsystemic approach (Cardey, 2013) and the Contextual Exploration Method (Desclés, 2006). The latter makes use of expressions called indicators, and of sets of linguistic clues to disambiguate the occurrences of the indicators. In the current paper we will only concentrate on the construction of a set of strong indicators of uncertainty in scientific texts that are defined as follows:
Definition: *A strong indicator of uncertainty is a linguistic expression, which can be composed of one or several lexical items contiguous or not, and which has the following properties: 1) it carries the semantic meaning of uncertainty; 2) (almost) all the sentences that contain the indicator express uncertainty, i.e. the indicator is very unlikely to appear in other sentences.*
To construct the A-set of strong indicators of uncertainty we have proceeded in the following way:

1. We have manually extracted of a set of 100 sentences that express uncertainty from the two datasets.

2. we extracted the possible strong indicators of uncertainty that are present in each sentence of this set (see examples in italic in table 4). These strong indicators were added to the A-set.

3. For each cue of the H-set and the C-set, we have considered expressions that include the cue, such that their presence in a sentence suffices to classify it as expression of uncertainty.

As a result, the A-set of strong indicators that we have obtained contains 20 expressions that can be presented as sequences of words and operators. Table 5 shows several examples of strong indicators[5]. Optional elements are indicated in brackets and alternatives are separated using "/". As some of the sequences can be non contiguous, this is expressed by an ellipsis.

2.4. Distributional analysis of uncertainty cues and strong indicators

Our goal is to collect several types of data related to uncertainty cues and indicators: their frequencies and positions in the different sections of articles. To do this, we performed the following steps:

1. Classification of sections by analysing section titles;

2. Sentence segmentation of all sections;

3. Identification of occurrences of the H-set, C-set and A-set in the sentences.

The first step, classification of sections, is done in order to indentify sections that belong to the four major types in the IMRaD structure (Introduction, Methods, Results and Discussion), as well as other section types that may occur in the datasets. The IMRaD structure is a rethorical framework for scientific papers that has been adopted in many areas, including the biomedical field, and several previous studies analysed different properties of this structure in the PLOS dataset, see, e.g., (Bertin et al., 2015; Bertin and Atanassova, 2017; Atanassova et al., 2016; Atanassova and Bertin, 2016).
We classified section titles into the following 7 categories: "Introduction", "Methods", "Results", "Discussion", "Background", "Conclusion" and "Supplementary

[5]The entire A-set will be published on DataCite (https://www.datacite.org/).

PubMed ID	Sentence
PMC5034003	Nanoparticles (NPs) have dimensions from 1 up to 100nm by common *consensus*.
PMC3397131	After convergence of the iterations, a '*consensus* grid' is calculated by taking the mean of all transformed individual configurations.
PMC3664788	These were selected as putative wild species-specific markers and were assembled using the CAP3 software, yielding seven assembled *consensus* sequences comprising 20 sequences in total.
PMC4666279	Stakeholders (drawn from industry and policy communities) have identified applications in the agri-food sector as being the potentially most *controversial* as far as societal acceptance is concerned (Gupta et al. 2013; Matin et al. 2012).
PMC5306339	Both panels use the same color code to describe the resulting classification: yellow for NM, green for non-NM, red and blue for an *inconsistent* classification between the two techniques.
PMC4047867	Our data are not *inconsistent* with these previous findings, but rather add further insight into FoxA1 function by suggesting a regulatory pathway mediated by FoxA1 and Tip30 in events controlling the expansion of ER+ luminal cells and ER+ mammary luminal tumor development.
PMC4886164	While this conclusion was initially *confusing* since a substantial body of earlier work shows that PI3K is central to the control of Na+ transport [...], the simplest explanation of these findings is that the basal level of PI3K activity is sufficient to ensure phosphorylation of SGK1.
PMC3920033	7 also indicate that the presence of ferritin (or its mimetics) in an *unknown* solution can be confirmed down to the relatively low iron concentrations of the order of 10-3-10-4g/L.
PMC4515301	It is not *surprising* that genes involved in the flavonoid biosynthesis are differentially expressed between these two cultivars because NAT produces white flowers and CAB blooms are red.

Table 3: Examples of sentences that contain cues from the C-set and do not express uncertainty.

PubMed ID	Sentence
PMC4260744	The present work *raises some doubts about the widely accepted* antioxidant potential of RSV.
PMC5376413	Still, *there is no evidence of* adipocyte dedifferentiation in vivo and *more studies are needed to* understand the process.
PMC4465843	These results *may enforce the concept that* these untranslated regions are prone to a higher level of environmental and evolutionary constraints compared to the coding sequences and *it is plausible that* selection shapes these lengths.
PMC5034002	*This is believed to be* related to the fact that the saturation concentration for Cu in solution was rapidly reached for the given exposure setting.
PMC5555889	Thus, *it is difficult to draw a general conclusion* based on studies that have used different methods and assessed different aspects of physical activity.
PMC3890610	*The most probable explanation is* mismatching of ionic and atomic radius of europium and yttrium appearing in strong lattice strength influenced by this difference.
PMC4300398	The results obtained again *seem to comply with* the LMD mechanism of particle formation, provided that the limited solubility of Ag-acetate in EG, and hence a need for its dissolution in the reaction system, is taken into account in addition to high reactivity.
PMC5555887	*This apparently strange result might be partly due to* less intake of rat chow and an alternate possibility is debated in the Discussion part.

Table 4: Examples of sentences that do express uncertainty using complex expressions.

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

Table 5: Examples of strong indicators of uncertainty from the A-set.

material". We used regular expressions that were constructed manually to capture the different variations that can exist in section titles, e.g., the "Methods" section can have titles such as "Materials and Methods", "Method", etc. The sentence segmentation was done by analysing the punctuation and capitalisation patterns in the text. We use sentences as major units to measure text progression.
Finally, to perform distributional analyses, the occurrences of the H-set, C-set and A-set were identified in all sentences, case-insensitive.

3. Results

In this section we present the results of the experimentation with the two datasets.

3.1. Section structure of the articles

After analysing the totality of the 46 395 section titles, about 81.89 % of the sections were classified into one of the 7 categories mentioned above. The remaining 8 404 sections correspond often to additional information sections such as "Conflict of Interest", "Taxonomy", "Disclosure Statement", "Authors' contributions", etc. Of course, some of the sections have titles that are domain specific and they could not be classified by our method.
Table 6 shows the most frequent article structures that we observed in the datasets. Section classes are expressed as follows: I - Introduction, M - Methods, R - Results, D - Discussion, B - Background, C - Conclusion, S - Supplementary material, X - unknown. RD stands for "Results and Discussion" when these two sections are merged.

Biomed dataset		Physics dataset	
Structure	Percentage	Structure	Percentage
R-D-M	26,29 %	I-X+-C	23,36 %
I-M-R-D-X$^+$	9,72 %	I-X$^+$-R-X	13,11 %
M-R-D-X$^+$	5,34 %	I-X$^+$	12,70 %
I-M-X-S	5,17 %	I-X$^+$-R-C	11,68 %
I-M-R-D-C-X	4,61 %	I-X$^+$-RD-C	7,17 %
I-M-R-D	4,37 %	I-M-RD-C	3,07 %
I-M-R-D-C	4,07 %	I-X-RD-C-S	2,66 %
B-M-R-D-C	3,83 %	I-M-RD-C-S	2,46 %
M-R-D	3,51 %	I-M-R-D-C-S	2,46 %
I-M-X-D-S	2,87 %	I-M-R-D-C	1,84 %

Table 6: Most frequent articles structures in the datasets.

We observe that the sections in the Biomed dataset display more regularities and were better classified by our method. Also, the majority of articles in the Biomed dataset share the sequence "I-M-R-D" or similar. Many of the sections in the Physics dataset could not be classified, and that is expressed by "X" in the table. Most of these sections, which occur in the middle of the papers, have titles that are domain specific. The rest of the papers in this dataset tend to follow the "I-M-R-D" sequence.

3.2. Distributions of uncertainty cues and indicators

Figure 1 describes the distributions of the H-set and the C-set along the text progression of the articles. The overall number of occurrences for the two sets is different: the H-set has 110 065 occurrences in the Biomed dataset and 7 833 occurrences in the Physics dataset, while the C-set has 15 715 and 6 364 occurresnces respectively. On figure 1 we observe that the C-set has distributions which vary considerably between the two datasets. In Physics, there is a clear peak at around 75 % of the text progression. One possible reason may be that this position roughly corresponds to the beginning of the Discussion or Conclusion section. In the Biomed dataset, the C-set has a peak in the beginning of the articles, most often the Introduction section.
Figure 2 presents the relative number of occurrences of the H-set and the C-set in the different section types. Here, we observe that both the H-set and the C-set have numerous occurrences in the Discussion section in the Biomed dataset. However, the occurrences of the C-set are significantly smaller in number in the identified categories. The last category "X" stands for sections that were not classified that are very frequent in the Physics dataset. For this reason, the relative number of occurrences of both sets is high for these sections.
Considering the A-set of strong indicators, 23 709 occurrences were identified in the Biomed dataset, and only 974 occurrences were identified in the Physics dataset. Figure 3 describes the distributions along the text progression. The occurrences become more frequent towards the end of the articles. The data for the Physics dataset is not sufficient to draw a general conclusion.
The distributions of the A-set and the H-set in the Biomed dataset are similar to each other. However, it must be noted that the number of occurrences for these two sets are very different: 23 709 for the A-set and 110 065 for the H-set. This difference comes from the fact that the H-set contains verbs that can be frequently used and that appear in many sentences that do or do not express uncertainty. The A-set, on the other hand, contains indicators that have fewer occurrences but all the sentences where they occur express uncertainty. Censequently, using A-set indicators over H-set cues results in much less noise in the identification of sentences with uncertainty.

4. Discussion

We have studied two different sets of uncertainty cues and proposed a new A-set of strong indicators of uncertainty. The results show that cues and strong indicators of uncertainty are most frequent in the Discussion section of articles and in general they tend to occur towards the end of the text progression. Significant differences can be observed between the two datasets that are most likely due to the fact that the predominant structure of the articles differs between the Biomedical dataset and the Physics dataset.
The A-set that is constructed in this way may not be exhaustive. The work presented here is a first step in the study of uncertainty cues and their positions in the IMRaD structure. Studying of expression of uncetainty in science can be further envisaged along three different axes: increasing the volume of the dataset that is processed, studying the variations that exist between the different disciplines, and finally more fine-grained analysis taking into consideration various degrees of uncertainty. Furthermore, creating a set

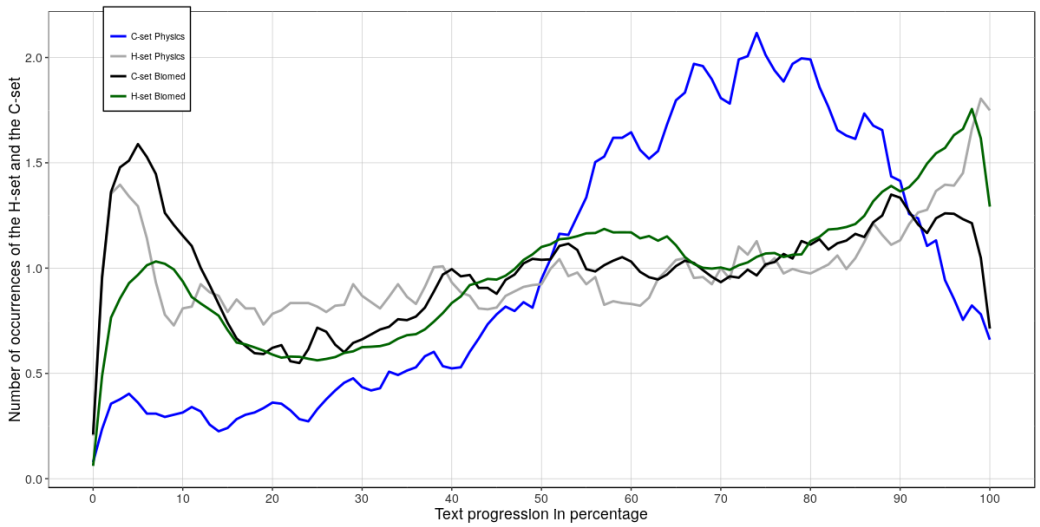

Figure 1: Distribution of the H-set and the C-set along the text progression.

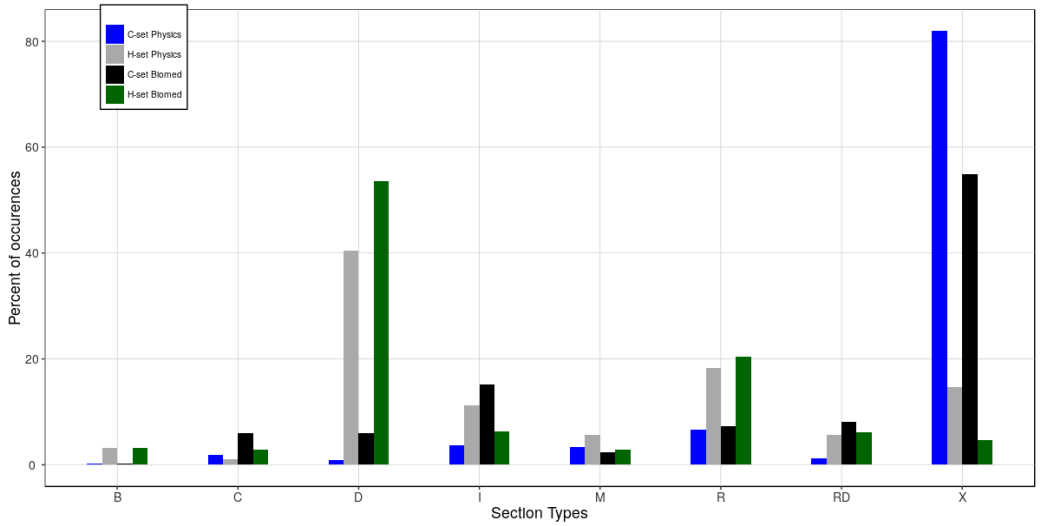

Figure 2: Occurrences of the H-set and the C-set with respect to section types.

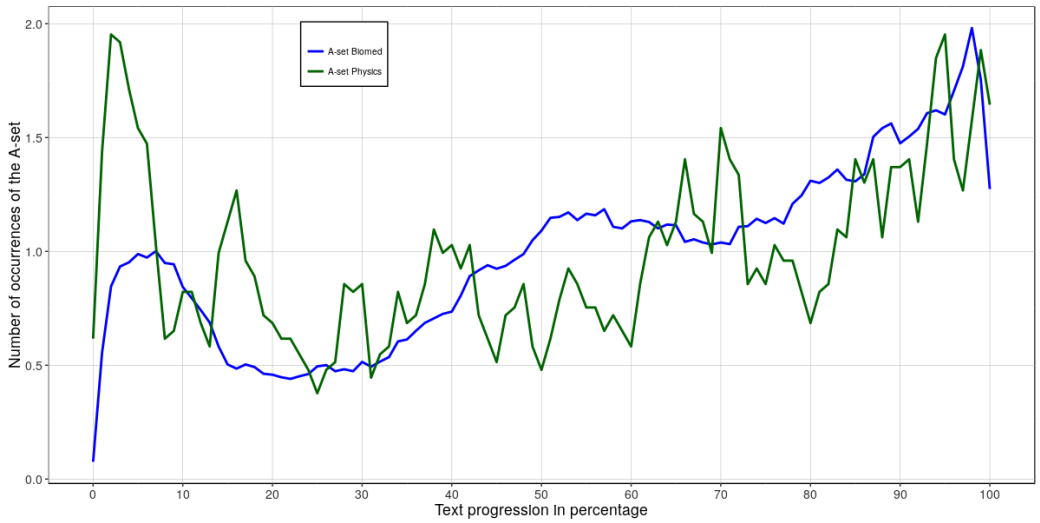

Figure 3: Distribution of the A-set along the text progression.

of strong indicators, such as the A-set, that can be qualified as exhaustive is a more difficult task and is not the objective of this study. For this reason, further evaluations need to be made to quantify the capacity of such a set to identify sentences that express uncertainty in an exhaustive way, which is generally given by the recall measure. The enrichment of the A-set to obtain better coverage can be envisaged by considering synonym expressions and manual annotations. The A-set is defined so as to ensure that the extraction of sentences which express uncertainty produces little or no noise. While this property needs to be further evaluated, we have focused in this study on the distributions of the cues and strong indicators, and on the structural properties of articles, rather than on proposing a method for the reliable extraction of sentences expressing uncertainty. The fact that few occurrences of the A-set were identified in the Physics dataset indicates that in the domains of Physics authors make use of different sets of expressions to convey the idea of uncertainty. This phenomenon will be studied in our future research.

While larger datasets exist, e.g. PubMed and arXiv, the utilisation of the IMRaD structure in articles strongly depends on the discipline. We plan to extend this study to other domains and examine the differences that exist between the disciplines.

5. Bibliographical References

Atanassova, I. and Bertin, M. (2016). Temporal properties of recurring in-text references. *D-Lib Magazine*, 9/10(22).

Atanassova, I., Bertin, M., and Larivière, V. (2016). On the composition of scientific abstracts. *Journal of Documentation*, 72(4):636–647.

Bally, C. (1965). *Linguistique generale et linguistique francaise: 4e ed. rev. et corrigee*. Berne.

Bernhard, D. and Ligozat, A.-L. (2011). Analyse automatique de la modalité et du niveau de certitude: application au domaine médical. *Actes de TALN*, 1:433–444.

Bertin, M. and Atanassova, I. (2017). The context of multiple in-text references and their signification. *International Journal on Digital Libraries*, Jul.

Bertin, M., Atanassova, I., Larivière, V., and Gingras, Y. (2015). The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology (JASIST)*.

Caers, J. (2011). *Modeling Uncertainty in the Earth Sciences*. John Wiley & Sons Ltd, Oxford, United Kingdom.

Cardey, S. (2013). *Modelling Language*. John Benjamins Publishing Company, Amsterdam / Philadelphia.

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

Committee, A. M. (1995). Uncertainty of measurement: implications of its use in analytical science. *Analyst*, 120:2303–2308.

Desclés, J., Alrahabi, M., and Desclés, J.-P. (2009). Bioexcom: Automatic annotation and categorization of speculative sentences in biological literature by a contextual exploration processing. In *Proceedings of the 4th Language & Technology Conference*, pages 32–40.

Desclés, J.-P. (2006). Contextual exploration processing for discourse and automatic annotations of texts. In *FLAIRS Conference*, volume 281, page 284.

Desclés, J., Makkaoui, O., and Hacène, T. (2014). biomedical texts : new perspectives and large- negation and speculation in natural language processing. In *Negation and Speculation in Natural Language Processing (NeSp-NLP 2010)*, number July 2010, pages 32–40, Uppsala, Sweden.

Farkas, R., Vincze, V., Móra, G., Csirik, J., and Szarvas, G. (2010). The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12. Association for Computational Linguistics.

Flemming, D., Feinkohl, I., Cress, U., and Kimmerle, J. (2015). Individual Uncertainty and the Uncertainty of Science: The Impact of Perceived Conflict and General Self-Efficacy on the Perception of Tentativeness and Credibility of Scientific Information. *Frontiers in psychology*, 6:1859.

Friedman, S., Dunwoody, S., and Rogers, C. (1999). Communicating uncertainty: Media coverage of new and controversial science.

Fusco, G., Bertoncello, F., Candau, J., Emsellem, K., Huet, T., Longhi, C., Poinat, S., Primon, J.-L., and Rinaudo, C. (2015). Faire science avec l'incertitude : réflexions sur la production des connaissances en Sciences Humaines et Sociales. In *Incertitude et connaissances en SHS : production, diffusion, transfert*, page 17, Nice, France. Maison des Sciences de l'Homme et de la Société Sud-Est (MSHS) - Axe 4 : Territoires, systèmes techniques et usages sociaux.

HAS - Haute Autorité de la Santé. (2013). Etat des lieux - Niveau de preuve et gradation des recommandations de bonne pratique. Technical report, HAS - Haute Autorité de la Santé, Saint-Denis La Plaine, France.

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

Hyland, K. (1996b). Writing without conviction? hedging in science research articles. *Applied Linguistics*, 17(4):433–454.

Hyland, K. (1998). *Hedging in scientific research articles*, volume 54. John Benjamins Publishing.

Johnson, B. B. and Slovic, P. (1995). Presenting uncertainty in health risk assessment: Initial studies of its effects on risk perception and trust. *Risk Analysis*, 15(4):485–494.

Joint Committee for Guides in Metrology. (2008). GUM - Évaluation des données de mesure - Guide pour l'expression de l'incertitude de mesure. Technical report, BIPM - Bureau International des Poids et Mesures, Sèvres, France.

Joint Committee for Guides in Metrology. (2012). VIM - Vocabulaire international de métrologie - Concepts fondamentaux et généraux et termes associés. Technical Re-

port VIM, 3e édition, BIPM - Bureau International des Poids et Mesures, Sèvres, France.

Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508.

Milliken, F. J. (1987). Three types of perceived uncertainty about the environment: State, effect, and response uncertainty. *The Academy of Management Review*, 12(1):133–143.

Moncecchi, G., Minel, J.-L., and Wonsever, D. (2012). Improving speculative language detection using linguistic knowledge. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 37–46. Association for Computational Linguistics.

Pfeffer, J., Salancik, G. R., and Leblebici, H. (1976). The effect of uncertainty on the use of social influence in organizational decision making. *Administrative Science Quarterly*, 21(2):227–245.

Shackley, S. and Wynne, B. (1996). Representing uncertainty in global climate change science and policy: Boundary-ordering devices and authority. *Science, Technology, & Human Values*, 21(3):275–302.

Weiss, C. (2003). Scientific uncertainty and science-based precaution. *International Environmental Agreements*, 3(2):137–166, Jun.

World Commission on the Ethics of Scientific Knowledge and Technology (COMEST). (2005). The Precautionary Principle. Technical report, UNESCO, Paris, France.

Wynne, B. (1992). Uncertainty and environmental learning: Reconceiving science and policy in the preventive paradigm. *Global Environmental Change*, 2(2):111–127.