https://doi.org/10.51408/issi2025_002

Is There Life on Mars? Studying the Context of Uncertainty in Astrobiology

Iana Atanassova[1], Panggih Kusuma Ningrum[2], Nicolas Gutehrlé[3], Francis Lareau[4], Christophe Malaterre[5]

[1]*iana.atanassova@univ-fcomte.fr*
Université Marie et Louis Pasteur, CRIT, F-25000 Besançon (France)
Institut Universitaire de France (IUF), Paris (France)

[2]*panggih_kusuma.ningrum@univ-fcomte.fr*
Université Marie et Louis Pasteur, CRIT, F-25000 Besançon (France)

[3]*nicolas.gutehrle@univ-fcomte.fr*
Université Marie et Louis Pasteur, CRIT, F-25000 Besançon (France)

[4]*lareau.francis@courrier.uqam.ca*
Université de Sherbrooke, Dept of Philosophy and Applied Ethics, Sherbrooke (QC) J1K 2R1 (Canada)
Université du Québec à Montréal, Dept of Philosophy & CIRST, Montréal (QC) H3C 3P8 (Canada)

[5]*malaterre.christophe@uqam.ca*
Université du Québec à Montréal, Dept of Philosophy & CIRST, Montréal (QC) H3C 3P8 (Canada)

Abstract

While science is often portrayed as producing reliable knowledge, scientists tend to express caution about their claims, acknowledging nuances and doubt, all the more so in novel domains of research paved with unknowns. Uncertainty is an intrinsic aspect of scientific inquiry, particularly in recent fields such as astrobiology, which tackles numerous hard questions about the origin, evolution, and distribution of life on Earth and elsewhere. Mapping uncertainty in science matters for achieving a more accurate understanding of scientific knowledge. It also helps identify research domains at the frontiers of knowledge where unknowns are the most salient. In this article, we investigate the presence, distribution and context of uncertainty in the field of astrobiology. We analyze a comprehensive corpus of 3,698 research articles published in three major journals in the domain from 1968 to 2020. We use a linguistically motivated approach to identify expression of uncertainty in article full text. The corpus was further segmented into research topics using Latent Dirichlet Allocation (LDA) to investigate variations in uncertainty across subfields and over time. Our findings show that, while uncertainty has remained relatively stable over the 50 years covered by the corpus, constituting 20–25% of sentences on average, it varies significantly across research fields, highlighting areas where unknowns, doubts and speculations are more prevalent. The analysis also highlights relationships between expression of uncertainty and rhetorical structure. Indeed, higher uncertainty levels were observed in the beginning (introductions) and towards the end (conclusions) of research articles, while middle sections contained less uncertainty. Abstracts also tended to express a slightly higher level of uncertainty compared to main texts, especially with greater variability, suggesting their role in summarizing research and highlighting unknowns. To investigate the context of uncertainty, a lexical analysis was conducted to identify nouns most frequently associated with uncertainty within each topic. Terms such as “life,” “planet,” and “Mars” were found to be strongly associated with uncertainty. Conversely, terms related to experimentation and measurement, such as “sample” and “spectrum,” were linked to an absence of uncertainty, pointing at a dichotomy between speculative and evidence-based lines of inquiry. The findings contribute to a better understanding of

the field of astrobiology and exemplify the relevance of the proposed method to identify uncertainty-related concepts in corpora of full text publications. They also offer a foundation for future comparative studies across disciplines.

Introduction

Uncertainty is a foundational element of scientific inquiry, influencing every stage of the research process from formulating hypotheses to interpreting results. The construction of new scientific knowledge, by its nature, involves various degrees of uncertainty, arising from research hypotheses, methodological limitations, measurement errors, and the interpretative nature of scientific reasoning. Therefore, studying uncertainty is important to gain understanding on the mechanisms behind the construction of new knowledge. It also matters for better depicting the status of scientific knowledge and its variations in evidential support in different fields of inquiry. Indeed, scientific fields vary not just in terms of objects of investigation but also in terms of methods, maturity of research programs and social organization, thereby likely displaying noticeable nuances in terms of uncertainty. In the present contribution, we propose to investigate how uncertainty is expressed in the recent discipline of astrobiology.

Astrobiology is a multidisciplinary field encompassing areas such as prebiotic chemistry, systems chemistry, synthetic biology, atmospheric sciences, planetary sciences, and astronomy that emerged in the 1990s following early works in space life sciences and origin of life studies (Dick & Strick, 2004). Its unifying feature is the pursuit of hard and yet unresolved questions that require cross-disciplinary insights: What is life? How did it originate on Earth? Does it exist elsewhere in the universe? How might life evolve on a cosmic scale? According to the NASA Astrobiology Roadmap, astrobiology includes the search for habitable exoplanets, Mars exploration, studies of life's origins and early evolution, and research on life's adaptability on Earth and in space (Des Marais et al., 2003). Similarly, the AstRoMap European Astrobiology Roadmap frames astrobiology as the study of life's origin, evolution, and distribution within cosmic evolution, addressing habitability in the Solar System and beyond (Horneck et al., 2016). The broad scope of astrobiology, as well as its recent emergence and the relatively speculative nature of its research objectives make it a perfect target for assessing the expressions of uncertainty in scientific research.

To this aim, we propose to deploy a linguistically motivated approach for identifying and categorizing uncertainty onto a full-text corpus consisting of all research articles published in the three major astrobiology journals (from earliest publication date in 1968 up until 2020). This approach relies on the identification of specific terminological patterns in texts, thereby going beyond more classical analyses of uncertainty focusing on hedgers and boosters (Ningrum & Atanassova, 2024). Moreover, by using a topic model already fitted to the corpus (Malaterre & Lareau, 2023), the method makes it possible to investigate uncertainty over time and across different subfields of astrobiology, thereby revealing nuances across disciplinary contexts which are further examined by identifying discriminating terms associated

with uncertainty. Uncertainty is also analyzed as a function of document properties and rhetorical structure (e.g., text progression, length, abstracts vs. main texts).
In what follows, we first describe the corpus and the methods, and lay out the set of analyses we conducted. We then present the results and discuss them, notably considering directions for future work.

Corpus and Methods

The corpus consists of all full-text research articles of the three major astrobiology journals that had been assembled in (Malaterre & Lareau, 2023): *Astrobiology*, the *International Journal of Astrobiology (IJA)*, and *Origins of Life and Evolution of Biospheres (OLEB*, this latter journal being successively known as *Space Life Sciences* (1968-1973), *Origins of Life* (1974-1984), *Origins of Life and Evolution of the Biosphere* (1984-2004), and *Origins of Life and Evolution of Biospheres* (2005-2023); since 2024, the journal has been renamed *Discover Life*). Editorials, conference summaries, errata, discussion notes, and short articles (<4,000 characters) were removed so as to only keep research articles and their abstracts. This led to a corpus consisting of a total of 3,698 full-text articles, including 3,542 with abstracts, from 1968 to 2020, with a total of 705,636 sentences (out of which 26,355 correspond to abstracts and 679,281 to the main text of the articles).
The corpus underwent standard preprocessing, including cleaning, tokenization, and vectorization. For the topic model, part-of-speech (POS) tagging and lemmatization using the TreeTagger package (Schmid, 1994) were conducted, and only nouns, verbs, modals, adjectives, adverbs, proper nouns, and foreign words were retained. Stop words, words shorter than three characters, and those appearing in fewer than 20 documents were removed. A topic model with K=25 topics was applied to the text data using the LDA algorithm (Blei et al., 2003), following a manual review of models with various K values (Malaterre & Lareau, 2023). Topics were interpreted and named based on an examination of top words and top texts. To facilitate analysis, the topics were organized into clusters using Louvain community detection on a graph of topic-to-topic correlations in documents. In short, the topics can be grouped into four clusters: (A) focuses on life and survival, including microbial communities in extreme environments, space biology, spacecraft contamination, and conceptual studies like Fermi's paradox. (B) centers on the origins of life, exploring prebiotic chemistry, amino- and nucleic acids, molecular evolution, meteorite analyses, and definitions of life, including artificial life and protocells. (C) addresses planetary and astro-related topics, such as exoplanet habitability, planetary atmospheres, chirality, and energy-matter delivery from space. (D) investigates biosignatures and geological traces, covering Mars exploration, hydrothermal vents, biopaleontology, microfossils, and the search for water and habitability on other worlds.
This study adopts a linguistically motivated system developed by Ningrum et al. (2023) to detect scientific uncertainty in scholarly full texts that is built using the spaCy framework. The system was applied to the cleaned full-text corpus (including abstracts). The system uses a weakly supervised approach with a fine-grained annotation scheme to identify uncertainty expressions at the sentence level. Its pipeline integrates pattern matching, complex sentence analysis, and authorial

reference checks, leveraging a span-based method to pinpoint uncertainty in academic writings. For a detailed presentation of this system, see Ningrum & Atanassova (2023). Building on prior findings (Desclés et al., 2011; Ningrum et al., 2025) that emphasize the importance of multi-word phrases in identifying hedging and uncertainty, the system goes beyond simple linguistic markers, and also relies on linguistic patterns and features, such as part-of-speech (POS) tags, morphology, and syntactic dependencies. Unlike earlier studies that assume all uncertainty expressions must contain at least one uncertainty span (Medlock & Briscoe, 2007; Szarvas, 2008; Farkas et al., 2010), this approach treats uncertainty spans as trigger candidates that require further verification. The verification covers three main types of contextual shifts that can alter the true interpretation of scientific uncertainty expression: rebuttal expressions due to confirmation, rebuttal expressions due to neutral informative statements, and negation. Figure 1 shows several examples of sentences and annotations. Table 1 presents a description of the dataset with the number of documents for each topic, the total number of sentences and the number of sentences identified as containing uncertainty. We processed abstracts and main texts of articles separately.

1 - *"Evaluation* ***seems to be an unresolved matter*** *in…."* [Uncertainty]

2 - *"The* ***potential roles of X*** *in Y* ***remain speculative****."* [Uncertainty]

3 – *"...no evidence to support* ***this hypothesis****..."* [Absence of uncertainty due to negation]

4 – *"In order to test* ***whether X has a contribution to Y****, statistical analysis was employed...."*

[Absence of uncertainty with neutral informative statement]

5 – *"The high correlations scores confirm* ***hypothesis H3*****"**

[Absence of uncertainty due to confirmation]

Figure 1. Examples of sentences and annotations of uncertainty. In bold: expressions of uncertainty that trigger the analysis of the context, and underlined: contextual elements that are analyzed to confirm or refute the presence of uncertainty.

Table 1. Dataset description for the abstracts and article main texts: number of documents, total number of sentences and sentences containing uncertainty for each topic. Articles were assigned their dominant topic as determined by LDA.

Dominant topic	Abstracts			Article main texts		
	Nb of documents	Total nb of sentences	Nb of sent. with uncertainty	Nb of documents	Total nb of sentences	Nb of sent. with uncertainty
A-Bacteria-microbes	178	1,560	312	179	32,980	5,146
A-Cell-plant-animal	110	851	127	118	20,128	3,454
A-Life-civilization	156	970	363	177	33,305	10,636
A-Radiation-spore	178	1,490	228	178	29,373	4,018
A-Sample-mission	83	659	99	87	25,718	4,754
A-Science-mission	72	523	61	86	17,508	3,059
B-Amino-acid	91	539	152	95	14,188	3,215
B-Chemistry	282	1,713	378	291	37,655	6,963
B-Life-system	241	1,531	462	256	42,069	11,315
B-Organic-molecule	232	1,658	461	238	39,586	8,858
B-Protein-gene-RNA	129	941	272	136	20,194	5,096
B-Sample-chemistry	204	1,505	303	211	32,200	5,612
B-Surface-mineral-vesicle	153	1,143	292	156	26,208	5,300
C-Atmosphere	123	1,090	349	128	32,524	8,681
C-Chirality	126	691	184	140	18,376	4,382
C-Impact-particle	107	804	279	107	21,510	5,885
C-Planet-star	156	1,260	407	160	36,779	9,533
C-Value-model	124	889	273	124	23,493	6,056
D-Life-environment	110	844	274	125	28,392	8,970
D-Mars	97	797	233	101	24,169	6,344
D-Reaction-vents	134	1,058	312	145	24,535	6,167
D-Rock-sample	104	904	268	106	23,093	5,600
D-Spectra	125	1,018	195	125	24,501	4,145
D-Structure-geology	149	1,292	288	151	32,494	6,710
D-Water	78	625	222	78	18,303	5,188
Total	3,542	26,355	6,794	3,698	679,281	155,087

Analyses

Once identified, sentences with uncertainty were summed up for each article and analyzed across the corpus, especially to assess the influence of time and research domains (topics). Three main uncertainty measures were calculated: uncertainty as a function of time period U_p; uncertainty as a function of topic and time period $U_{j,p}$; and uncertainty as a function of topics U_j:

$$U_p = \frac{\sum_{d \in p} u_d / s_d}{N_p} \qquad U_{j,p} = \frac{\sum_{d \in p} u_d \times t_{j,d}}{\sum_{d \in p} s_d \times t_{j,d}} \qquad U_j = \frac{\sum_p U_{j,p}}{T}$$

where u_d is the number of sentences expressing uncertainty in a document d, s_d is the number of sentences in document d, N_p is the number of documents per time period p, $t_{j,d}$ is the % value of topic j in document d, T is the number of time periods (18 in the present case). This was done for abstracts only, for main texts only, and

for complete articles (abstracts and main texts jointly) to compare uncertainty expressed in abstracts and main texts.

Further analyses were conducted to examine uncertainty as a function of text length and text progression (excluding abstracts in both cases), the latter being defined as:

$$\text{for a given } g \in \{0, 1, \ldots, 100\},\ U_g = \frac{\sum_d u_{d,g}}{\sum_d s_{d,g}}$$

where $u_{d,g}$ is the number of sentences expressing uncertainty such that their relative position $h = rank(s) \times 100/s_d$, where $rank(s)$ is the absolute position of sentence s in document d (from 1 to s_d), is such that g is the entire number that is closest to h.

To investigate the context of uncertainty in astrobiology, we analyzed occurrences of nouns and proper nouns in the body of the articles in each identified topic. First, we extracted the most frequent nouns from sentences annotated with uncertainty for each topic, thus identifying key terms that frequently occur around expressions of uncertainty. To do this, we performed tokenization, POS-tagging and lemmatization of the dataset using the Python Natural Language Toolkit (NLTK). Articles were assigned their dominant topic as determined by the LDA topic model. Second, we calculated Precision, Recall, and F-measure scores for these nouns to assess their effectiveness in characterizing uncertainty. The F-measure is a metric used to evaluate the performance of a classification model, particularly in information retrieval and machine learning (Van Rijsbergen, 1979; Christen et al., 2024). In the context of classification and feature selection, it has been shown that the F-measure can be used to rank features with respect to their degree of association with a class (e.g., Alwidian et al., 2016; Lamirel et al., 2016). With this in mind, for a given term *t* and a set *S* of all the sentences of a given set *D* of documents, we define a class-association score $A_{c,t,S}$ that expresses the degree of association of *t* with a given class *c* in *S* as the harmonic mean:

$$A_{c,t,S} = 2 \times \frac{AP_{c,t,S} \times AR_{c,t,S}}{AP_{c,t,S} + AR_{c,t,S}}$$

where

$$AP_{c,t,S} = \frac{|\{s:\, s \in S \cap c, t \in s\}|}{|\{s:\, s \in S, t \in s\}|} \text{ and } AR_{c,t,S} = \frac{|\{s:\, s \in S \cap c, t \in s\}|}{|\{s:\, s \in S \cap c\}|},$$

s is a sentence, and class *c* is a class that can be either "presence of uncertainty" or "absence of uncertainty". This approach enabled us to identify and rank the nouns which were most strongly associated with the presence (or absence) of uncertainty within each topic, in order to better understand the primary subjects or concepts related to uncertainty discourse across the different topics in the dataset. We specifically examined this context for the top 5 and bottom 5 topics with respect to U_j. We also calculated class-association scores for the nouns of the top 5% and the bottom 5% of the articles (i.e.,u_d/s_d), in order to identify frequent concepts associated with uncertainty.

Results

The main results are summarized hereafter, with contrasting variations in uncertainty depending on context, notably research topics, but also rhetorical dimensions such as text length and text progression.

Uncertainty as a function of time

Results indicate a relatively stable expression of uncertainty over the fifty year span of the corpus, in the range of about 20 to 25% of document sentences (abstracts and main texts together) (Fig. 2). Percentage of uncertainty sentences can be as low as about 5%, while maximum uncertainty may reach about 50%, with some outlier documents scoring even above 60%. In any case, most of the corpus documents express a relatively high level of uncertainty which remains relatively unchanged over time, despite the introduction of two new journals in 2000 and underlying changes in topics (Malaterre & Lareau, 2023).

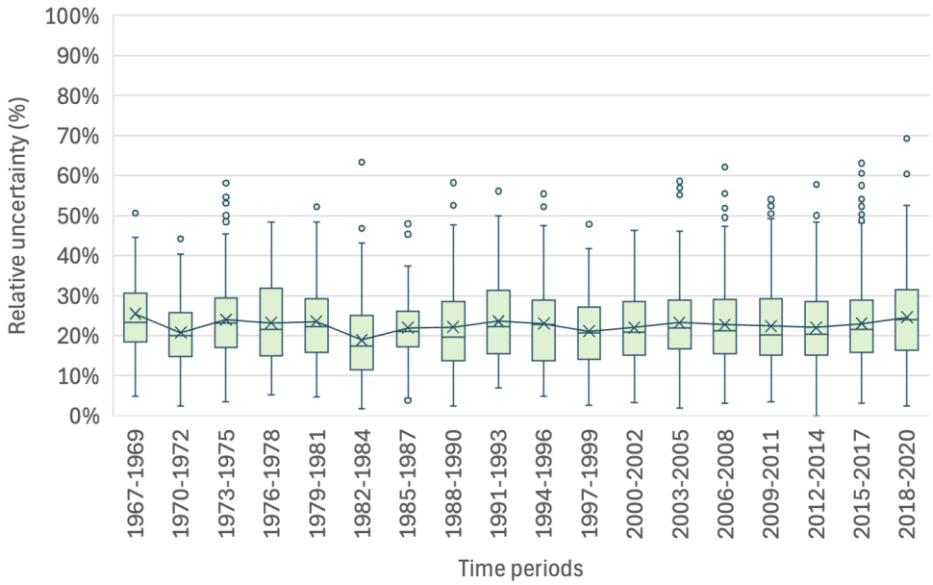

Figure 2. Evolution of share of uncertainty sentences in articles over time. For each time-period, boxplot of the distribution statistics of uncertainty percentage in research articles (abstract and main text); the line represents the evolution of average uncertainty U_p; dots are outlier articles.

Uncertainty per topic (excluding abstracts)

Analysis of uncertainty as a function of topic shows significant variation: while some topics express uncertainty in as few as about 15% of their attributed sentences, other topics have their share of uncertainty sentences well above 25% (Fig. 3). Among the five topics with least uncertainty, one finds three topics related to space microbiology ("A-Radiation-spore", "A-Bacteria-microbes", "A-Cell-plant-animal"), one to chemical analysis of rock samples ("B-Sample-chemistry), and one related to spectral analyses ("D-Spectra"). Among the five topics with the most uncertainty,

three concern life, its environment, whether alien civilization exists, what it means for a system to be alive ("D-Life-environment", "A-Life-Civilization", "B-Life-System") and two that concern astronomy, planetary systems in particular and impactors ("C-Planet-star", "C-Impact-Particle").

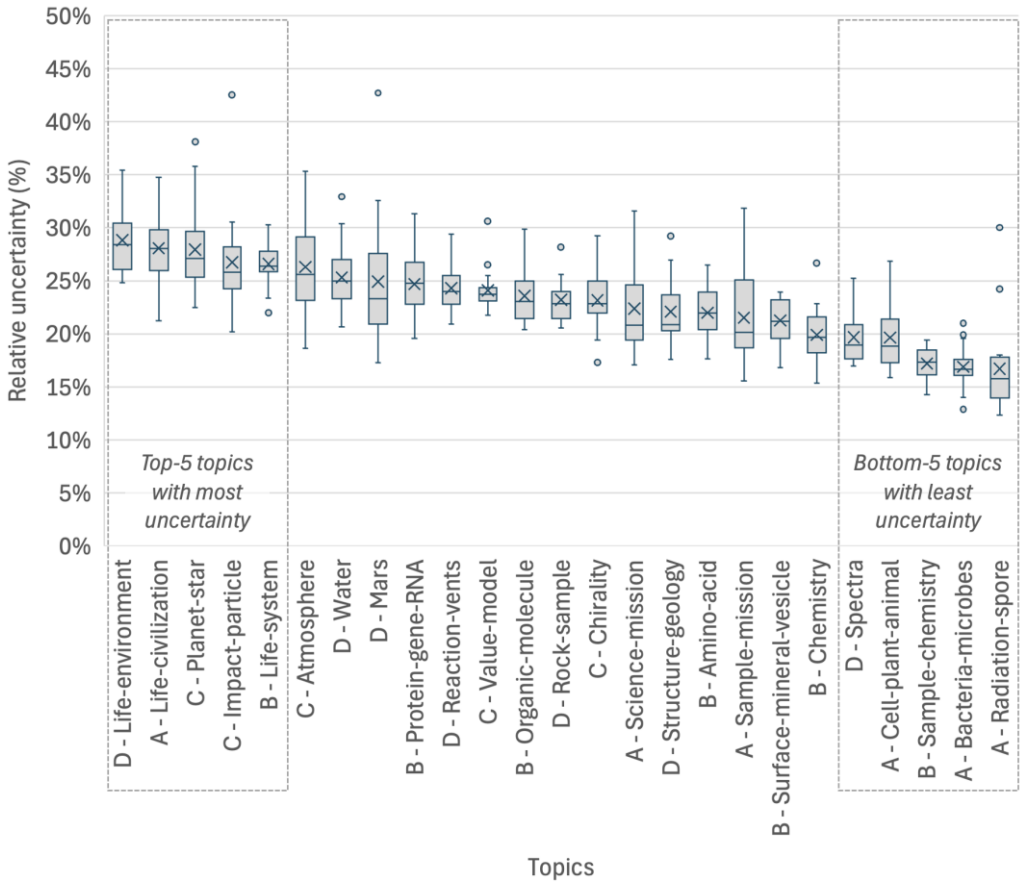

Figure 3. Share of uncertainty sentences as a function of topics. For each topic, boxplot of the distribution statistics of uncertainty % $U_{j,p}$ attributed to each of the 25 topics (for sentences in the main text only as abstract were not included in the topic model) across the 18 time-periods of the study; dots are outlier time-periods.

Context of uncertainty

To better understand the contexts of uncertainty, we examined the association scores with "uncertainty" and "absence of uncertainty" of all the nouns appearing in the corpus for various sets of documents (topic-related documents, outliers, and all corpus documents). Tables 2 and 3 present two different aspects of this analysis. As the different topics in the dataset contain various degrees of relative uncertainty (see Fig. 3), and the same phenomenon can be observed at the article level, we analyzed these association scores to identify the concepts that are most commonly related to the presence of uncertainty or to its absence. The highest association scores we observed on the dataset are about 0.36, thus the scores vary between 0 and 0.36. At

the article level, relative uncertainty in the main text varies between 1.43% and 69.81%. Due to this large interval, for the following analysis, we examined the outliers defined as the top 5% and the bottom 5% of articles, and compared them to the top-5 topics in terms of uncertainty, the bottom-5 topics, and to all corpus articles together.

Table 2 summarizes the highest association scores—above 0.1—with "presence of uncertainty" for the top-5 topics, top 5%articles, and all articles. We can observe, for example, that the noun "life" is highly related to the expression of uncertainty across all 5 topics, being at top position for three of them (D-Life-environment, A-Life-civilization, B-Life-system), and within the top-5 terms for the other two topics (C-Planet-star and C-Impact-particle). "Life" is also the highest ranking term among the top 5% articles expressing the most uncertainty, and across all articles of the corpus, but to a lesser extent. The nouns "planet" and "Earth" are present in all lists except for one topic (B-Life-system). Each topic presents its specificities, e.g. the uncertainty in D-Life-environment is prominently related to objects such as "environment", "Mars", "surface" and "condition" that do not appear in the other lists. Similarly, B-Life-system expresses uncertainty related to "molecule", "evolution" and "process" which are specific to that topic.

Table 3 lists the nouns that exhibit the highest association scores with "absence of uncertainty". We calculated these scores for the 5 topics that have the lowest relative uncertainty, for the bottom 5% articles in terms of relative uncertainty, and for all articles. Here, the term "sample" appears on the first or the second position for all lists except one topic (A-Cell-plant-animal). The lists that were obtained for the bottom 5% of articles and for all articles contain only one term ("sample") and no terms respectively. This can be explained by the much higher number of sentences without uncertainty compared to the number of sentences with uncertainty (about 4-fold, see Table 1); hence a much more diverse set of statements and vocabulary that cannot have high association scores with any specific noun.

Comparison between Tables 2 and 3 underscores insights on the types of research objects that are related to uncertainty within the different topics. Several terms in Table 3 appear related to experimentation and evidence-based research, e.g. "spectrum", "sample", "band", "cell", "experiment", "study", "acid", "temperature", "solution", "reaction", "spore". These nouns are strongly associated with the absence of uncertainty. In contrast, Table 2 indicates that uncertainty is expressed in relation with objects more prone to speculation or objects that are less directly observable or amenable to experimentation, e.g. "life", "planet", "Mars", "civilization", "star", "system", "atmosphere", "water", "evolution". Additionally, some objects can be related to the absence of uncertainty in some domains (e.g., "water" in the topic B-Sample-chemistry in Table 3), while being associated with the presence of uncertainty in other topics (D-Life-environment and C-Planet-star in Table 2). The term "time" is related to uncertainty for 3 topics in Table 2 but does not appear in Table 3. Similarly, "life" is prominently associated with uncertainty, while being absent from Table 3.

Table 2. Nouns with the highest association scores (above 0.1) with uncertainty for: the top 5 topics with highest relative uncertainty; the top 5% of articles with highest relative uncertainty; and all articles. Association scores are given in parentheses.

Topics with highest relative uncertainty										Top 5 % of articles		All articles	
D-Life-environment		A-Life-civilization		C-Planet-star		C-Impact-particle		B-Life-system					
life	(0.357)	life	(0.241)	planet	(0.352)	impact	(0.200)	life	(0.199)	life	(0.253)	life	(0.151)
Earth	(0.216)	planet	(0.122)	star	(0.212)	Earth	(0.178)	system	(0.188)	Earth	(0.164)	Earth	(0.116)
environment	(0.161)	Earth	(0.116)	Earth	(0.153)	life	(0.130)	molecule	(0.122)	planet	(0.136)	surface	(0.107)
planet	(0.160)	civilization	(0.110)	system	(0.152)	time	(0.127)	evolution	(0.119)	surface	(0.123)	planet	(0.101)
Mars	(0.139)	time	(0.106)	life	(0.132)	event	(0.122)	process	(0.109)	time	(0.116)		
surface	(0.129)			mass	(0.121)	planet	(0.109)			water	(0.112)		
water	(0.111)			atmosphere	(0.113)								
condition	(0.109)			time	(0.105)								
				water	(0.103)								

Table 3. Nouns with highest association scores (above 0.1) with *absence of uncertainty* for: the bottom 5 topics with lowest relative uncertainty; the bottom 5% of articles with lowest relative uncertainty; and all articles. Association scores are given in parentheses.

Topics with lowest relative uncertainty										Bottom 5 % of articles		All articles
D-Spectra		A-Cell-plant-animal		B-Sample-chemistry		A-Bacteria-microbes		A-Radiation-spore				
spectrum	(0.224)	cell	(0.183)	acid	(0.224)	sample	(0.194)	sample	(0.224)	sample	(0.141)	no terms
sample	(0.180)	experiment	(0.104)	sample	(0.200)	cell	(0.110)	radiation	(0.177)			
Raman	(0.153)	study	(0.103)	experiment	(0.140)			cell	(0.170)			
band	(0.136)			water	(0.111)			condition	(0.146)			
surface	(0.100)			solution	(0.107)			space	(0.142)			
				temperature	(0.106)			spore	(0.142)			
				reaction	(0.101)			experiment	(0.139)			
				compound	(0.101)			exposure	(0.134)			
								temperature	(0.107)			

Uncertainty in abstracts and in main texts

Uncertainty expressed in abstracts and in the body of articles tend to follow the same relatively stable pattern over time, though uncertainty in abstracts is usually a few points above uncertainty in the core of the texts. Note the higher variability of uncertainty expressed in abstracts, with most abstracts oscillating between 10% and 40% uncertainty, with minima at 0% and maxima or outliers oscillating between 80% and 100% uncertainty in some cases. The spread of uncertainty in the body of articles is much narrower, typically in between 15% and 25% of sentences expressing uncertainty.

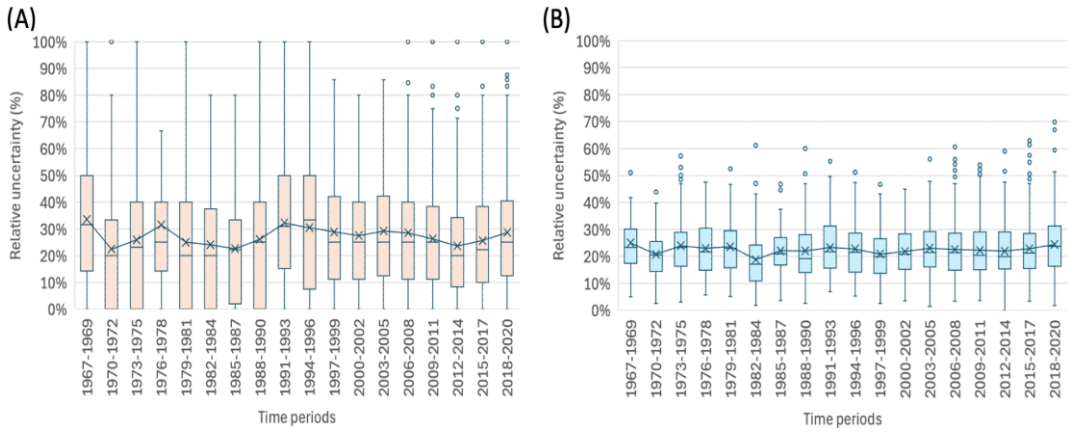

Figure 4. Comparison of uncertainty expressed in abstracts (A) and in the main portion of the corpus articles (B). Boxplot showing the distribution of document uncertainty ratio; line representing the evolution of average uncertainty per time-period.

Uncertainty as a function of text length

Analyzing text length as a function of uncertainty shows a lot of variability, though a noticeable trend seems to indicate that texts with either low or high uncertainty tend to be on the short side (around 100 sentences for texts with less than 10% uncertainty or more than 55%). On the other hand, texts with average uncertainty tend to be longer (about 200 sentences for texts with 20-30% uncertainty). This suggests that polarized texts in terms of uncertainty, exhibiting either a lot of doubt or a lot of conviction, tend to be on the shorter end.

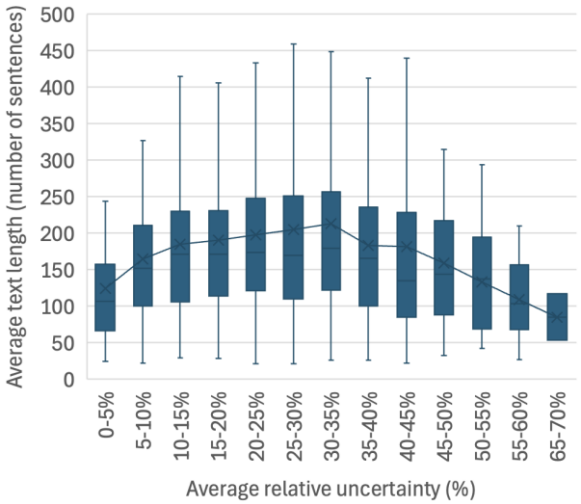

Figure 5. Document length as a function of uncertainty. For different intervals of uncertainty percentage in documents, boxplot of the distribution statistics of corresponding document length (total number of sentences in abstracts and main texts jointly).

Uncertainty as a function of text progression

Figure 6 shows the plot of the percentage of sentences that express uncertainty U_g with respect to their position in the text progression. Far from being constant throughout a text, uncertainty significantly fluctuates depending on text progression.

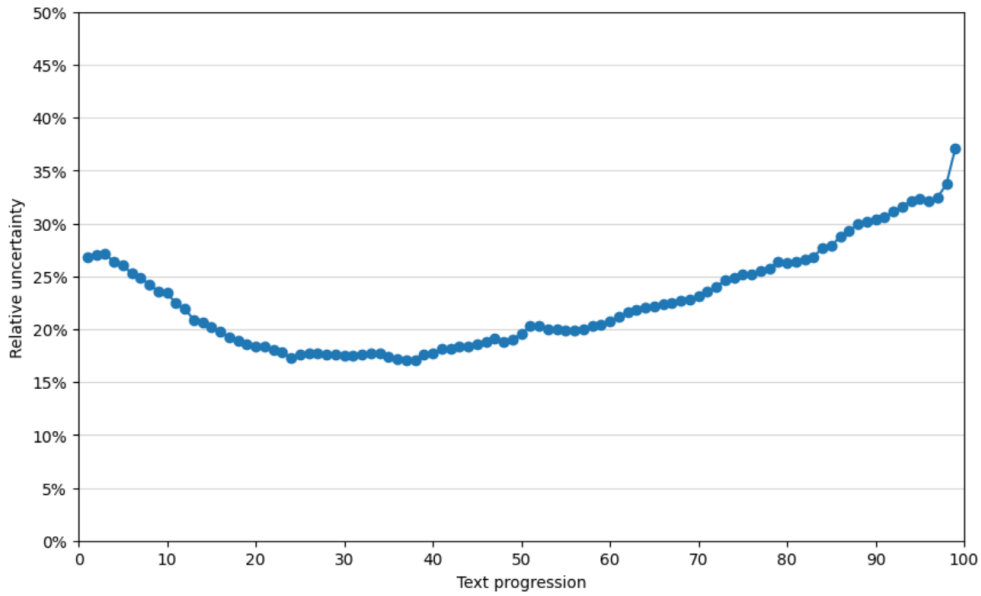

Figure 6. Relative distribution of uncertainty as a function of text progression (for main texts only, abstract excluded; 0 corresponds to text start; 100 to end).

The introductory portions of texts display a relatively high share of uncertainty, with as many as 27% of sentences expressing uncertainty. An even higher level of uncertainty is expressed in the concluding sections, with average uncertainty up to 37% at the end of texts. In between these two extremes, uncertainty levels are lowest between positions 20 and 40 of text progression. The IMRaD (Introduction, Methods, Results and Discussion) structure for articles is most usual in experimental sciences and commonly used in the journals in our dataset. Assuming such a structure for the majority of the corpus articles, uncertainty levels are rather high in the Introduction of the articles, at their lowest around the middle of the texts, i.e. in the Method and Result sections, and increase towards the final Discussion section.

Discussion

Our approach to annotating uncertainty, while effective, is not without limitations. The annotation relies on a set of nuanced linguistic rules to identify uncertainty, yielding an F-measure of 0.858 (Ningrum et al., 2025). While this performance is robust, it is not perfect and may introduce noise. Recent methodological improvements have been made, and further enhancements are planned.

The topic modeling approach which was used to identify research domains also has its constraints. We employed Latent Dirichlet Allocation (LDA) as it represents a well-established method, and fitted the model to K=25 topics so as to offer a

reasonable balance between granularity and research objectives, as addressed in prior work (Malaterre & Lareau, 2023). While providing nuanced topic probability distributions, this approach can impose certain limitations, for instance, the need for additionally crisp-assigning documents by assigning them to their dominant topics in some analyses.

Finally, the dataset itself presents limitations as it is confined to specific journals and time periods, reflecting a disciplinary focus on astrobiology. While this focus aligns with our objective of investigating uncertainty in that specific nascent multidisciplinary domain, extending the corpus to include articles from other journals using keyword-based retrieval could provide broader insights.

Our findings reveal that uncertainty in astrobiology research articles is relatively stable over time, both across the entire corpus and within specific topics. Contrary to initial expectations, uncertainty did not decrease over time, even as the field matured. While this challenges the hypothesis that uncertainty diminishes with disciplinary maturation, it remains possible that this trend could emerge at finer topic granularity than the 25 topics used in this study.

The corpus demonstrates relatively high levels of uncertainty, with on average about 20-25% of sentences in articles expressing uncertainty. This contrasts with previous studies that reported an average of 14% uncertainty in corpora from generalist and biomedical journals (Ningrum & Atanassova, 2024). Astrobiology thus occupies the higher end of the spectrum in terms of expressed uncertainty in the corpora examined so far. Moreover, individual articles vary widely, with some exhibiting as much as 60% uncertainty and others less than 10%. Investigating these extreme cases could yield valuable insights into the factors driving such variability.

One major finding is the significant variability in uncertainty across research topics. Certain topics express markedly more uncertainty, often linked to specific objects of inquiry. For example, particular nouns frequently associated with uncertainty suggest that the nature of the research object influences the level of expressed uncertainty. In the present study, the F-measure was used to identify most strongly associated nouns with specific groups of documents, yet the scores are low and furthermore the data is unbalanced; other measures, such as micro F-measure or TF-IDF at the cluster level (Grootendorst, 2022), could be used in future works. Future investigations should also explore in more detail whether epistemic properties—such as the difficulty of experimentation, observational challenges, or complexity—underlie this variability. One direction is to investigate the relationships between uncertainty and specific epistemic markers as defined in (Malaterre & Léonard, 2024). Additional sociological or cultural factors, such as differences in writing styles or practices, may also contribute and warrant further study.

Our analyses also highlight the interplay between uncertainty and the rhetorical structure of research articles. While there is no significant difference in average uncertainty between abstracts and main texts, abstracts exhibit greater variability in uncertainty levels. Interestingly, shorter texts tend to polarize in terms of uncertainty, displaying either very high or very low levels. Text progression emerges as a major variable influencing uncertainty. The introduction, discussion, and conclusion sections account for most instances of uncertainty, suggesting that these sections

function as rhetorical spaces for articulating doubt, speculation, and reflection. Comparative analyses across the IMRaD structure in different fields could further elucidate these patterns.

Conclusion

Deploying a linguistically motivated approach to identify complex terminological patterns expressing uncertainty in scientific articles, this study highlights the intricate dynamics of uncertainty within astrobiology research, offering insights into its relative stability over time, its variability across subdomains of research, and different facets of its rhetorical manifestations. Despite the field's maturation over the past fifty years, uncertainty remains prevalent, reflecting the challenges of investigating the origin on Earth and its possible presence elsewhere in the solar system and beyond. The variability of uncertainty across research domains—as captured with topic modeling —underscores different regimes of uncertainty possibly linked to specific objects of enquiry and their properties, and which will need to be further investigated. Lexical analysis identified nouns frequently linked to uncertainty, such as "life," "planet," and "Mars," contrasting with terms like "sample" and "spectrum," which reveal evidence-based inquiry. The analyses also highlight the relationship between uncertainty and the rhetorical structure of scientific articles. Higher uncertainty is found in introductions and conclusions, while middle sections contain less. Abstracts show slightly higher and more variable uncertainty, emphasizing their role in summarizing research and unknowns. These findings not only contribute to our understanding of the science of astrobiology and the uncertainty that pervades it, but also open pathways for comparative studies with other corpora and methodological refinements, notably to identify different types of uncertainties and further examine the epistemic context in which uncertainty is expressed. By extending these lines of enquiry, future research can further illuminate the nuanced role of uncertainty in scientific discourse.

Acknowledgments

C.M. acknowledges funding from Canada Social Sciences and Humanities Research Council (Grant 430-2018-00899) and Canada Research Chairs (CRC-950-230795). F.L. acknowledges funding from Canada Social Sciences and Humanities Research Council (Postdoctoral Fellowships 756-2024-0557). I.A., N.G. and P.K.N. acknowledge funding from the French ANR Project InSciM "Modelling Uncertainty in Science" (2021-2025) under grant number ANR-21-CE38-0003-01.

References

Alwidian, J., Hammo, B., & Obeid, N. (2016). Enhanced CBA algorithm based on apriori optimization and statistical ranking measure. *Proceeding of 28th International Business Information Management Association (IBIMA) Conference on Vision, 2020*, 4291–4306.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Christen, P., Hand, D. J., & Kirielle, N. (2024). A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Computing Surveys*, *56*(3), 1–24. https://doi.org/10.1145/3606367

Des Marais, D. J., Allamandola, L. J., Benner, S. A., Boss, A. P., Deamer, D., Falkowski, P. G., Farmer, J. D., Hedges, S. B., Jakosky, B. M., Knoll, A. H., Liskowsky, D. R., Meadows, V. S., Meyer, M. A., Pilcher, C. B., Nealson, K. H., Spormann, A. M., Trent, J. D., Turner, W. W., Woolf, N. J., & Yorke, H. W. (2003). The NASA Astrobiology Roadmap. *Astrobiology*, *3*(2), 219–235. https://doi.org/10.1089/153110703769016299

Desclés, J., Alrahabi, M., & Desclés, J.-P. (2011). BioExcom: Detection and Categorization of Speculative Sentences in Biomedical Literature. In Z. Vetulani (Ed.), *Human Language Technology. Challenges for Computer Science and Linguistics* (pp. 478–489). Springer. https://doi.org/10.1007/978-3-642-20095-3_44

Dick, S. J., & Strick, J. E. (2004). *The Living Universe NASA and the Development of Astrobiology*. Piscataway, NJ: Rutgers University Press.

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

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Horneck, G., Walter, N., Westall, F., Grenfell, J. L., Martin, W. F., Gomez, F., Leuko, S., Lee, N., Onofri, S., Tsiganis, K., Saladino, R., Pilat-Lohinger, E., Palomba, E., Harrison, J., Rull, F., Muller, C., Strazzulla, G., Brucato, J. R., Rettberg, P., & Capria, M. T. (2016). AstRoMap European Astrobiology Roadmap. *Astrobiology*, *16*(3), 201–243. https://doi.org/10.1089/ast.2015.1441

Lamirel, J.-C., Dugué, N., & Cuxac, P. (2016). New efficient clustering quality indexes. *2016 International Joint Conference on Neural Networks (IJCNN)*, 3649–3657.

Malaterre, C., & Lareau, F. (2023). The Emergence of Astrobiology: A Topic-Modeling Perspective. *Astrobiology*, *23*(5), 496–512. https://doi.org/10.1089/ast.2022.0122

Malaterre, C., & Léonard, M. (2024). Epistemic Markers in the Scientific Discourse. *Philosophy of Science*, *91*(1), 151–174. https://doi.org/10.1017/psa.2023.97

Medlock, B., & Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 992–999.

Ningrum, P. K., & Atanassova, I. (2024). Annotation of scientific uncertainty using linguistic patterns. *Scientometrics*, *129*, 6261–6285. https://doi.org/10.1007/s11192-024-05009-z

Ningrum, P. K., Guterhlé, N., & Atanassova, I. (2025). Étudier l'incertitude dans les articles scientifiques: Mise en perspective d'une méthode linguistique. *Conférence Extraction et Gestion Des Connaissances (EGC)*.

Ningrum, P. K., Mayr, P., & Atanassova, I. (2023). UnScientify: Detecting Scientific Uncertainty in Scholarly Full Text. *Proceedings of ACM Conference (Workshop'23)*. EEKE2023, Santa Fe, New Mexico, USA.

Schmid, H. (1994). Part-of-speech tagging with neural networks. *Proceedings of the 15th Conference on Computational Linguistics-Volume 1*, 172–176. https://doi.org/10.3115/991886.991915

Szarvas, G. (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. *Proceedings of Acl-08: HLT*, 281–289.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths and Co.