Etudier l'incertitude dans les articles scientifiques : mise en perspective d'une méthode linguistique

Panggih Kusuma Ningrum*, Nicolas Gutehrlé*,
Iana Atanassova*,**

*Université de Franche-Comté, CRIT, F-25000 Besançon, France
panggih_kusuma.ningrum@univ-fcomte.fr
nicolas.gutehrle@univ-fcomte.fr
**Institut Universitaire de France (IUF), France
iana.atanassova@univ-fcomte.fr

Résumé. L'incertitude fait partie intégrante du processus de recherche scientifique et est inhérente à la construction de nouvelles connaissances. Dans cet article, nous examinons la manière dont l'incertitude est exprimée dans les articles scientifiques et proposons un cadre d'annotation rendant compte des différentes dimensions de cette notion. L'incertitude scientifique est définie ici comme l'expression d'un manque de connaissance ou d'un manque de précision dans les informations sur un sujet ou un concept identifié. Nous proposons un jeu de données de référence (*gold standard*), composé de 1 839 phrases d'articles scientifiques annotées manuellement et provenant de plusieurs disciplines. Nous proposons également une approche à base de connaissances linguistiques pour l'annotation automatique des articles et pour la détection et la catégorisation de l'incertitude scientifique. Nous comparons l'efficacité de notre approche en termes de scores de Précision, Rappel et F1 aux méthodes de prompts *few-shot* réalisées via les Grands Modèles de Langue Phi-3.5 et Llama 3 pour la même tâche d'annotation. Cette évaluation comparative montre des scores similaires entre les approches, allant jusqu'à des scores F1 de 0,858 pour notre approche.

1 Introduction

L'incertitude en science fait partie intégrante du processus de recherche. Si la production de nouvelles connaissances s'appuie sur une démarche méthodologique rigoureuse en fonction des objets d'étude et des champs disciplinaires, la recherche engendre des incertitudes liées à l'utilisation d'outils ou observations ayant des marges d'erreur, ou bien à travers des raisonnements abductifs et inductifs. Dans ce travail, nous étudions l'incertitude scientifique à travers une modélisation ontologique et linguistique de cette notion, afin de proposer des méthodes pour son extraction et sa catégorisation dans les articles scientifiques. Nous définissons l'incertitude scientifique comme l'*expression d'un manque de connaissance ou d'un manque de précision dans les informations sur un sujet ou un concept identifié dans un texte scientifique*.

Dans cet article, nous proposons un cadre d'annotation rendant compte des différentes dimensions de l'incertitude scientifique et une méthode à base de connaissances linguistiques

pour son annotation. Nous proposons également un jeu de données de référence (*gold standard*), qui nous permet d'évaluer notre approche. Enfin, nous comparons ces résultats avec les performances de Grands Modèles de Langues (GMLs) sur la même tâche.

Le reste de l'article est organisé comme suit : nous présentons la notion d'incertitude, ainsi qu'un état de l'art de travaux portant sur cette notion, dans la Section 2. Dans la Section 3, nous présentons le corpus de phrases exprimant de l'incertitude que nous avons constitué à partir de revues scientifiques issues de diverses disciplines, ainsi que le guide d'annotation que nous avons suivi pour annoter ce corpus. Dans la Section 4, nous présentons notre approche pour la détection et la catégorisation de l'incertitude scientifique, qui repose sur un ensemble de connaissances linguistiques. Nous présentons le protocole pour évaluer les performances de notre approche, ainsi que les GMLs que nous avons sélectionnés, sur cette tâche dans la Section 5, avant de discuter des résultats obtenus par ces approches sur notre corpus dans la Section 6. Nous présentons enfin les limitations et la conclusion de ce travail, ainsi que des pistes de travaux futurs, dans les Sections 7 et 8 respectivement.

2 Etat de l'art

L'incertitude est un concept complexe, et la littérature propose plusieurs définitions du terme (voir par ex. Walker et al. (2003); Refsgaard et al. (2007); Ascough et al. (2008)). Sur le plan conceptuel, il convient de différentier l'incertitude de la subjectivité, tout en prenant en considération la modalité épistémique. L'incertitude est l'un des composants des résultats de la recherche scientifique, dont les dimensions sont multiples dans les différents champs disciplinaires. Les concepts et relations autour de la notion de modalité en linguistique ne sont pas consensuels (Halliday et Matthiessen, 2014).

L'identification des segments textuels exprimant une incertitude a été l'objectif de plusieurs travaux. Par exemple, l'identification de phrases spéculatives dans des textes par des approches en apprentissage automatique a été abordée par Moncecchi et al. (2012), qui souligne la spécificité de la problématique de la subjectivité. Un état de l'art dans ce domaine est présenté par Diaz et López (2019). L'étude de Chen et al. (2018) propose d'identifier les expressions d'introduction de l'incertitude en élargissant un ensemble restreint d'expressions. Cependant, ces derniers travaux expriment une vision binaire de l'incertitude, et n'abordent pas les différents niveaux et dimensions de l'incertitude afin de rendre compte de la complexité de cette notion. Pour aller plus loin, Atanassova et al. (2018) construisent une liste d'indicateurs forts de l'incertitude et observent les distributions de ceux-ci au sein de la structure rhétorique des articles en biomédecine et en physique.

La problématique de la compréhension interdisciplinaire et conceptuelle de la notion d'incertitude a été abordée par Rey et al. (2018), à travers l'étude d'un corpus d'articles scientifiques traitant du réchauffement climatique. Ils proposent un schéma relationnel de l'incertitude scientifique, où les incertitudes exprimées sont organisées sous forme de classes selon le type de raisonnement utilisé (abductif, inductif, déductif) et la présence ou non d'indications quantitatives sur l'incertitude. Cependant, identifier et catégoriser l'incertitude associée aux connaissances scientifiques à travers l'étude des contenus textuels demeure un défi. Le problème fondamental réside dans la difficulté de travailler avec des données textuelles présentes dans les articles scientifiques. La majorité des études précédentes se sont concentrées sur la détection et l'identification d'un ensemble spécifique d'indices ou de marqueurs d'incertitude

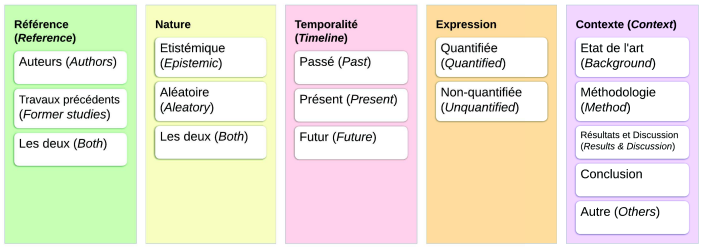

FIG. 1 – *Schéma d'annotation de l'incertitude. Les équivalents en anglais sont présentés en italique.*

dans les résumés d'articles scientifiques uniquement (Vincze et al., 2008; Guillaume et al., 2017) ou dans les articles en plein texte (Medlock et Briscoe, 2007; Atanassova et al., 2018; Riccioni et al., 2021). Bien que ces travaux aient contribué à l'élargissement du vocabulaire de l'incertitude, la mise en œuvre de ces techniques reste difficile en raison du travail manuel important qu'elles nécessitent et de la complexité du concept d'incertitude et de ses expressions en langage naturel.

3 Corpus annoté

Nous avons constitué un corpus interdisciplinaire de revues dans les domaines des Sciences Humaines et Sociales (SHS) et des Sciences, Technologies et Médecine (STM) afin d'étudier l'expression de l'incertitude dans les articles scientifiques. La sélection des revues de notre corpus s'appuie sur la classification *Scimago Journal and Country Rank* (SJR) [1], qui repose sur Scopus [2], la plus grande base de données académique disponible en ligne. Nous avons sélectionné des revues couvrant diverses disciplines, telles que la médecine, la biochimie, la génétique et la biologie moléculaire, l'informatique, les sciences sociales, les sciences de l'environnement, la psychologie, les arts et les sciences humaines. Nous avons sélectionné les cinq journaux les plus haut classés pour chaque discipline. De plus, nous avons inclus les revues PLoS ONE et Nature, toutes deux interdisciplinaires et haut classées.

Notre objectif est de produire un corpus annoté de phrases exprimant de l'incertitude scientifique. Pour cela, nous avons utilisé la catégorisation de l'incertitude scientifique en cinq dimensions proposée par Ningrum et Atanassova (2024). Chaque phrase a été annotée comme exprimant ou n'exprimant pas de l'incertitude (`Uncertainty` et `No Uncertainty`). Les phrases exprimant de l'incertitude ont ensuite été annotées selon cinq dimensions : Référence (*Reference*), Nature, Contexte (*Context*), Temporalité (*Timeline*), et Expression. La Figure 1 présente l'ensemble des étiquettes d'annotation utilisées, en français et an anglais.

A partir du corpus d'articles issus de différentes disciplines décrit ci-dessus, nous avons constitué un ensemble de phrases annotées de la manière suivante :

1. `https://www.scimagojr.com/journalrank.php`
2. `https://www.elsevier.com/products/scopus`

— 593 ont été pré-sélectionnées de manière automatique, par l'étude des occurrences des listes d'indices de l'incertitude proposées par Hyland (1996); Chen et al. (2018); Bongelli et al. (2019). Ces phrases ont été annotées manuellement par deux annotateurs.
— Le reste des phrases a été extrait à partir d'un sous-ensemble d'articles, composé de deux articles par revue sélectionnés aléatoirement. Ces articles ont été examinés par deux annotateurs pour y identifier et annoter les phrases porteuses d'incertitude.

Une description détaillée de ce processus est présenté dans le travail séminal de Ningrum et Atanassova (2024). Les annotateurs sont des étudiants de master Traitement Automatique des Langues en France, qui ont travaillé sur le corpus sur une période de 6 mois. Ces derniers ont été formés à partir d'un guide d'annotation et de phrases préalablement annotées afin de garantir la cohérence des annotations. Ils atteignent un score d'accord moyen de 0,414 selon le test Kappa de Cohen, montrant la difficulté de la tâche d'annotation de l'incertitude scientifique. Enfin, les annotations conflictuelles ont été résolues par un troisième annotateur indépendant.

Notre corpus final est ainsi composé d'un total de 1 839 phrases provenant de 445 articles de 21 revues en anglais, provenant de 8 disciplines différentes. La Table 1 présente la distribution d'articles et de phrases par revue dans notre corpus. La Figure 2 présente quant à elle la distribution des annotations selon les cinq dimensions décrites dans notre cadre d'annotation, tandis que des exemples de phrases annotées sont présentés dans la Table 2. Notre corpus est disponible sur Zenodo selon les principes de l'Open Science (Gutehrlé et al., 2024).

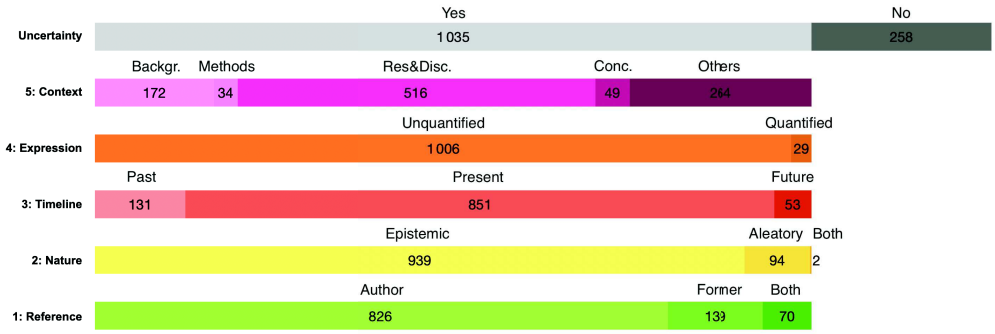

FIG. 2 – *Distribution des annotations pour chaque dimension de l'incertitude scientifique.*

4 Annotation basée sur des connaissances linguistiques

Nous proposons une approche à base de connaissances linguistiques pour identifier les expressions linguistiques d'incertitude dans les phrases. Notre approche emploie un ensemble de règles qui reposent sur la détection d'indices et marqueurs textuels (Vincze et al., 2008; Ningrum et Atanassova, 2024), ainsi que sur des patrons lexico-syntaxiques pour extraire les expressions d'incertitude dans les phrases. Une démo de notre approche est disponible en ligne [3].

Pour produire ces règles, nous avons tout d'abord identifié à partir des phrases de notre corpus les segments textuels exprimant de l'incertitude. Il convient de noter qu'une même phrase

3. https://bit.ly/unscientify-demo

Discipline	Revue	Nb. articles	Phrases U	Phrases No-U	Total phrases
Biochemistry, Genetics & Molecular Biology	Cell Reports Medicine	68	116	134	250
	Nature Communications	33	92	19	111
	Nucleic Acids Research	52	34	133	167
	Signal Transduction and Targeted Therapy	22	36	9	45
Computer Science	Applied Computing and Informatics	1	2	0	2
	Human-centric Computing and Inf. Sciences	2	10	1	11
Environmental Science	Environment International	2	41	1	42
	Perspective in Ecology and Conservation	3	12	3	15
Interdisciplinary	Nature	35	63	122	185
	PLoS One	43	24	129	153
Medicine	BMC Medicine	51	68	78	146
	Cardiovascular Diabetology	31	92	14	106
	Cell. and Mol. Gastroenterology and Hepatology	25	13	124	137
	Emerging Infectious Diseases	31	85	20	105
	Journal of Stroke	34	63	11	74
Psychology	Int. Journal of Clinical and Health Psychology	2	35	1	36
	Journal of Cognition	2	59	0	59
Social Science	Crime Science	2	40	0	40
	International Journal of STEM Education	2	17	0	17
	Sophia	2	103	5	108
Arts & Humanities	Music Theory Online	2	30	0	30
	Total	*445*	*1 035*	*804*	*1 839*

TAB. 1 – *Distribution d'articles et de phrases annotées par revue dans le corpus. U = "Uncertainty", Non-U = "No Uncertainty".*

Phrase	Annotation	Réference	Nature	Contexte	Temp.	Expr.
With these vectors, anti-cancer drugs can be delivered to tumors much more effectively than by circulatory delivery alone [23].						
	NO UNCERTAINTY					
In dilation, the estimation of the output pixel is the most extreme estimation of the considerable number of pixels in the input pixel's neighborhood.						
	UNCERTAINTY	AUTHOR	EPISTEMIC	METHODS	PRESENT	UNQUANT.
Some studies suggest that natural mutant PPARγ alleles might impair native PPARγ function [29] and differences in the PPARγ genotype could modify the response to TZD treatment.						
	UNCERTAINTY	FORMER	ALEATORY	RES&DISC	PAST	UNQUANT.
Contrary to our findings, a study in northern Kenya reported a high ZIKV seroprevalence of 7% and DENV seroprevalence of 1%, although it was not clear how the potential cross-reactivity was assessed [12].						
	UNCERTAINTY	BOTH	EPISTEMIC	RES&DISC	PAST	UNQUANT.

TAB. 2 – *Exemples de phrases issues de notre corpus et annotées selon notre cadre d'annotation.*

peut contenir plusieurs expressions d'incertitude. Nous avons constitué un premier ensemble de règles visant à déterminer si une phrase exprime de l'incertitude ou non, suivant les 14 sous-catégories d'incertitude définies par Ningrum et al. (2023). Nous avons également constitué un second ensemble de règles dites de résolution de référence, qui visent à déterminer la dimension Référence lorsqu'une phrase exprime de l'incertitude, c'est-à-dire à déterminer si cette incertitude est exprimée par les auteurs (`author`), par une étude précédente (`former studies`) ou par les deux (`both`). La Table 3 présente le nombre de règles de chaque type, tandis que la Figure 3 présente des exemples de segments textuels annotés. Le second exemple présente une phrase comprenant plusieurs expressions d'incertitude.

Règles de détection / rejet de l'incertitude				Règles de résolution de la Référence	
1-Explicite	28	8-Non-généralisable	3	Référence	22
2-Modalité	11	9-Adverbial	6		
3-Conditionnelle	2	10-Négation	4		
4-Hypothétique	8	11-Subjective	4		
5-Prédictive	1	12-Conjecturale	11		
6-Interrogative	7	13-Plausibilité	6		
7-Désaccord	4	14-Rejet	6		

TAB. 3 – *Distribution de règles pour la détection et le rejet de l'incertitude et de résolution de la Référence.*

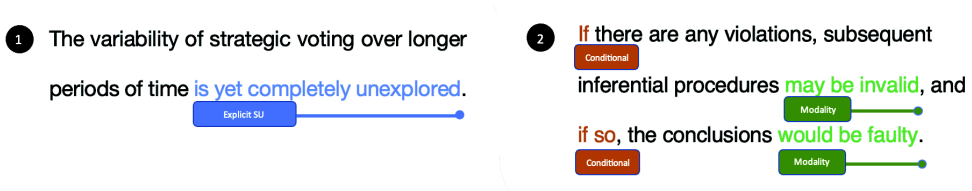

FIG. 3 – *Exemples de phrases exprimant divers types d'incertitude scientifique.*

L'implémentation de notre approche comprend quatre étapes : *pré-traitement*, *application des règles*, *résolution des rejets* et *résolution de la référence*, comme présenté dans la Figure 4. Nous employons le modèle `en_core_web_trf` mis à disposition par l'outil `spaCy` pour implémenter nos règles et réaliser les analyses en parties du discours, morphologique et en dépendances syntaxiques des phrases, qui sont nécessaires à notre approche.

L'étape de *pré-traitement* vise à nettoyer la phrase d'éléments textuels inutiles (par ex. sauts de lignes, espaces en fin de phrase, etc.) à l'aide d'expressions régulières. Nous remplaçons ensuite toutes les citations bibliographiques par une forme normalisée "`@CITATION`", qui est prise en compte dans les règles de résolution de Référence. Enfin, nous procédons à l'étape d'analyse linguistique de la phrase.

L'étape *application des règles* permet d'identifier les éléments textuels exprimant de l'incertitude dans la phrase. Il suffit qu'une seule de ces règles se déclenche pour passer à l'étape suivante de *résolution de rejet*. Cette étape vise à détecter la présence éventuelle d'expressions venant infirmer l'expression de l'incertitude dans la phrase. Dans le cas contraire, la phrase

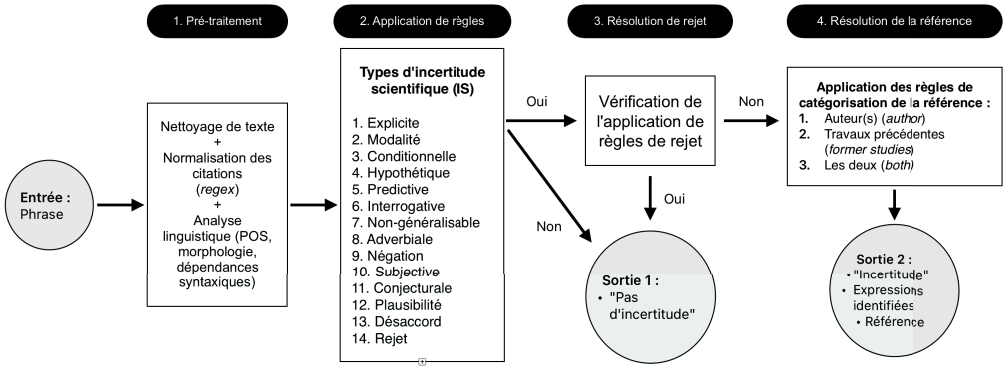

FIG. 4 – *Étapes principales de l'annotation par notre approche à base de connaissances linguistiques.*

est annotée comme "Incertitude" (`Uncertainty`). Enfin, pour les phrases annotées, nous appliquons l'étape *résolution de la référence*, qui annote la dimension Référence.

5 Protocole d'évaluation

L'évaluation de notre méthode se fait en deux étapes. Tout d'abord, nous évaluons les performances de notre algorithme d'annotation sur le corpus annoté présenté en Section 3, en terme de Précision, Rappel et F-mesure. Dans un deuxième temps, nous comparons notre approche aux performances de Grands Modèles de Langue (GML) existants. Pour cela, nous avons choisi les modèles Phi-3.5 (Abdin et al., 2024) et Llama 3 8B (Dubey et al., 2024) en raison de leur faible nombre de paramètres (3,8 et 8 milliards respectivement). Malgré ce nombre réduit, ces modèles peuvent atteindre des scores similaires voir supérieurs aux modèles plus grands sur diverses tâches impliquant du raisonnement sur des jeux de données standards. Ainsi, nous souhaitons observer les performances de modèles de petite et moyenne taille sur la tâche d'identification et de catégorisation de l'incertitude scientifique.

Nous employons un prompt identique pour soumettre la phrase à annoter aux GML. Dans ce prompt, nous demandons au modèle d'adopter la personnalité d'un chercheur, et de déterminer si la phrase exprime une incertitude ou non. Si c'est le cas, le modèle doit également indiquer si l'incertitude est exprimée par les auteurs de l'article, par des travaux précédents, ou les deux. Nous fournissons la définition de l'incertitude présentée en Section 1 au modèle, ainsi qu'un ensemble de cinq exemples de phrases annotées, sélectionnées manuellement depuis notre corpus afin d'illustrer les différentes annotations possibles. Les phrases employées comme exemples dans le prompt sont exclues du jeu de données.

Pour cette expérimentation, nous employons les modèles pré-entraînés mis à disposition par l'outil `Ollama`. Afin de limiter au maximum la variabilité des réponses générées par les modèles, nous fixons à 0 la valeur de l'hyper-paramètre `temperature`. Par ailleurs, étant donné qu'un modèle peut générer des réponses différentes à chaque lancement, les phrases ont été annotées trois fois par chaque modèle. Nous considérons l'annotation majoritaire parmi les trois réponses comme résultat du modèle.

6 Résultats

Les scores obtenus par les trois approches sur la tâche de détection de l'incertitude sont présentés dans la Table 4. Llama 3 atteint le meilleur score de Précision, qui est de 0,933, bien que les scores de Précision sont similaires entre les approches. A l'inverse, notre approche atteint des scores de Rappel et F1 de 0,812 et 0,858 respectivement, qui surpassent ceux obtenus par les deux autres approches.

	P	R	F1
Notre approche	0,909	**0,812**	**0,858**
Phi-3.5	0,912	0,525	0,666
Llama3	**0,933**	0,502	0,653

TAB. 4 – *Scores de Précision, Rappel et F1 obtenus par l'approche par connaissances linguistiques, Phi-3.5 et Llama 3 sur la tâche de détection de l'incertitude scientifique. Les meilleurs scores de Précision, Rappel et F1 sont présentés en gras.*

		P	R	F1
Notre approche	author	**0,954**	0,582	0,723
	former studies	0,362	0,911	0,518
Phi3.5	author	0,943	0,898	0,920
	former studies	0,624	0,857	0,722
Llama3	author	0,894	**0,953**	**0,923**
	former studies	0,470	0,893	0,616

TAB. 5 – *Scores de Précision, Rappel et F1 obtenus par l'approche linguistique, Phi-3.5 et Llama 3 sur la tâche de catégorisation de la source d'incertitude à l'incertitude scientifique. Les meilleurs scores sont présentés en gras.*

Pour la deuxième tâche d'annotation, qui est la catégorisation de la source d'incertitude en tant que "auteur" (`author`), "travaux précédents" (`former studies`) ou "les deux" (`both`), les résultats sont présentés dans la Table 5. Cette dernière table prend uniquement en compte les phrases correctement annotées en tant que Incertitude (`Uncertainty`). L'étiquette d'annotation `both` a été considérée comme une double annotation, c'est-à-dire à la fois `author` et `former studies`, pour cette évaluation. Les trois approches obtiennent des scores de Précision élevés pour la catégorie `author`. Llama 3 surpasse légèrement les autres approches, et atteint des scores de Rappel et F1 de 0,953 et 0,923 pour la catégorie `author`. Cependant, notre approche obtient un score de Précision très élevé de 0,954 pour cette catégorie, qui surpasse légèrement ceux obtenus par les autres approches. A l'inverse, les trois approches obtiennent des scores moins élevés pour la catégorie `former studies`.

Près de 66 % des erreurs produites par notre approche linguistique sur la tâche de détection de l'incertitude sont causées par neuf règles, tandis que près de 44 % des erreurs restantes ont été provoquées par des règles n'ayant été appliquées qu'une seule fois. De plus, trois phrases exprimant de l'incertitude ont été incorrectement rejetées par les règles de rejet. Sur la

tâche de catégorisation de la source d'incertitude, près de 20 % des phrases ont incorrectement été annotées comme `former studies` par seulement deux règles, tandis que trois règles sont responsables des phrases incorrectement annotées comme `both`. Ainsi, il serait possible d'améliorer ces résultats par l'amélioration d'un petit nombre de règles qui ont été identifiées.

L'analyse des réponses des GML montre que Llama 3 a généré trois réponses identiques pour chacune des phrases analysées. De plus, Llama 3 a généré uniquement les annotations demandées dans la majorité des cas, alors que seules 44 réponses générées comportent également une justification de la réponse en plein texte, qui ont nécessité un post-traitement pour les enlever. Phi-3.5 a systématiquement généré des réponses différentes lors des trois analyses des mêmes phrases, et ses réponses comportent toutes des justifications en plein texte. Les deux premières réponses sont identiques dans 79 % des cas. Par ailleurs, 42 % des réponses 1 et 3, et 55 % des réponses 2 et 3 sont identiques. Nous remarquons également qu'un même élément textuel peut servir à produire deux justifications contradictoires. En raison de ces variations, Phi-3.5 a produit huit réponses conflictuelles pour la tâche de détection de l'incertitude, et 13 réponses conflictuelles pour la tâche de catégorisation de la source d'incertitude. Ces conflits ont tous été résolus par vote majoritaire. Cependant, l'évaluation montre que la moitié de ces résolutions sont incorrectes. Cela montre que la fiabilité des GML est très variable.

7 Discussion et limitations

La méthode d'annotation à base de règles linguistiques proposée ici souffre des limitations habituelles liées à ce type d'approches, notamment la possibilité d'une certaine subjectivité dans la construction des règles et d'une plus grande difficulté de produire un rappel élevé. Cependant, les annotations produites permettent une traçabilité complète du processus d'annotation et ainsi la correction des erreurs est possible de manière incrémentale.

Le corpus multidisciplinaire que nous avons constitué reste limité : certains domaines scientifiques, qui ne sont pas présents dans notre corpus, font appel à l'incertitude de manière courante, tels que les domaines du *machine-learning* ou des *data sciences*, qui emploient des mesures de variabilité ou des tests d'hypothèse. Ces domaines feront l'objet d'études futures.

L'expérimentation avec les Grands Modèles de Langue que nous proposons sert ici d'un point de comparaison avec la méthode linguistique plutôt que d'une proposition d'une véritable méthode pour identifier l'incertitude. En effet, de nombreux paramètres pourraient être pris en compte pour rendre une telle expérimentation plus robuste. Nous avons fixé la valeur du paramètre `temperature` à 0, afin de limiter la variabilité des réponses générées. Cette valeur est responsable de la similarité dans les réponses générées par Llama 3 et Phi-3.5 lors de l'évaluation, et pourrait influencer la qualité des annotations obtenues. D'autres valeurs de `temperature` pourraient être prises en compte. De plus, nous avons utilisé un seul prompt, alors que l'expérimentation avec différentes formulations du prompt pourraient améliorer les performances, en particulier par des techniques de construction de prompts, telle que la technique de `chaîne de pensée` (*chain-of-thought*). Enfin, un entraînement des modèles avec des techniques *few shot learning* pourrait permettre aux modèles de mieux s'adapter à la tâche.

Dans ce travail, nous avons expérimenté avec les modèles Phi-3.5 et Llama 3, qui comportent respectivement 3,5 et 8 milliards de paramètres. Si les résultats de ces modèles sont déjà encourageants, il faudrait également évaluer les performances de modèles plus grands tels que Llama 70B, GPT-4 ou Mistral 8x22B sur la tâche de détection et de catégorisation

de l'incertitude. Cependant, les coûts énergétiques et d'entraînement pourraient également être considérés comme facteur dans la comparaison entre modèles et leurs performances. Notons que le coût en ressource et en énergie de notre approche linguistique est bien plus faible que n'importe quel modèle utilisant l'apprentissage profond, tout en permettant une analyse rapide.

8 Conclusion

Dans cet article, nous avons examiné la manière dont l'incertitude est exprimée dans les articles scientifiques. Nous avons tout d'abord proposé une définition de la notion d'incertitude en science, ainsi qu'un état de l'art des travaux autour de l'identification des incertitudes dans les articles. Nous avons ensuite présenté notre corpus annoté, constitué de 1 839 phrases issues d'articles scientifiques de diverses disciplines. Nous avons également précisé le cadre d'annotation que nous avons suivi pour annoter l'incertitude dans ces phrases avant de proposer une méthode pour l'identification et l'annotation des incertitudes dans des phrases, basée sur des connaissances linguistiques. Les performances de notre méthode ont été évaluées en termes de Précision, Rappel et F1, et en comparaison avec celles des Grands Modèles de Langues Phi-3.5 et Llama 3 8B sur la tâche de détection et de catégorisation de l'incertitude sur notre corpus. Cette évaluation montre que notre approche obtient les meilleurs scores de Rappel et F1 sur la tâche de détection de l'incertitude, tandis que le modèle Llama 3 obtient des scores supérieurs aux autres approches sur la tâche de catégorisation de la source d'incertitude.

La suite de ce travail s'orientera vers l'amélioration et l'extension de l'approche linguistique, afin de catégoriser les quatre autres dimensions de l'incertitude présentes dans le cadre d'annotation. Par ailleurs, la production de corpus annotés suffisamment grands pour envisager l'entraînement de nouveaux modèles de langues est également considérée.

Les applications de ce travail sont multiples. D'une part, le travail sur les modèles linguistiques peut permettre de produire de grands corpus annotés afin d'entraîner des modèles et rendre possible l'annotation à grande échelle des publications scientifiques. En parallèle, sur le plan épistémologique, l'étude de l'incertitude permettrait de révéler des différences entre les différentes disciplines, notamment par la comparaison entre revues, éditeurs scientifiques, et communautés, et d'identifier les différences entre les objets épistémiques auxquels se rapporte l'incertitude dans les différents domaines scientifiques. Dans le travail présenté ici, nous avons traité un corpus multidisciplinaire dans le but de refléter autant que possible la diversité des expressions de l'incertitude dans les disciplines scientifiques. Cependant, une approche comparative entre ces dernières pourrait apporter un nouvel éclairage sur les manières dont les chercheurs identifient et rapportent l'incertitude liée à leurs objets d'étude.

Remerciements

Ce travail est soutenu par le projet InSciM (2021–2025), financé par l'ANR, numéro de subvention ANR-21-CE38-0003-01. Les auteurs souhaitent également remercier vivement Marine Potier et Maya Mathie pour leur travail sur la constitution et l'annotation du corpus.

Références

Abdin, M., S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al. (2024). Phi-3 technical report : A highly capable language model locally on your phone. *arXiv preprint arXiv :2404.14219*.

Ascough, J., H. Maier, J. Ravalico, et M. Strudley (2008). Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological Modelling 219*(3), 383–399, doi: https://doi.org/10.1016/j.ecolmodel.2008.07.015. The Importance of Uncertainty and Sensitivity Analysis in Process-based Models of Carbon and Nitrogen Cycling in Terrestrial Ecosystems with Particular Emphasis on Forest Ecosystems.

Atanassova, I., F.-C. Rey, et M. Bertin (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.

Bongelli, R., I. Riccioni, R. Burro, et A. Zuczkowski (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, doi: 10.1371/journal.pone.0221933.

Chen, C., M. Song, et G. E. Heo (2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics 12*(1), 158–180, doi: 10.1016/j.joi.2017.12.004.

Diaz, N. P. C. et M. J. M. López (2019). *Negation and speculation detection*, Volume 13. John Benjamins Publishing Company.

Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv :2407.21783*.

Guillaume, J. H. A., C. Helgeson, S. Elsawah, A. J. Jakeman, et M. Kummu (2017). Toward best practice framing of uncertainty in scientific publications : A review of Water Resources Research abstracts. *AGU Publications Water Resources Research*, doi: 10.1002/2017WR020609.

Gutehrlé, N., P. K. Ningrum, et I. Atanassova (2024). Annotated Dataset for Uncertainty Mining : Gold Standard.

Halliday, M. A. K. et C. M. I. M. Matthiessen (2014). *Halliday's Introduction to Functional Grammar* (4$^{\text{th}}$ edition ed.). Routledge, Abingdon, United Kingdom.

Hyland, K. (1996). Talking to the Academy : Forms of Hedging in Science Research Articles. *Written Communication 13*(2), 251–281, doi: 10.1177/0741088396013002004. Publisher : SAGE Publications Inc.

Medlock, B. et T. Briscoe (2007). Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic. Association for Computational Linguistics.

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

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

Ningrum, P. K., P. Mayr, et I. Atanassova (2023). Unscientify : Detecting scientific uncertainty in scholarly full text. In *Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE- AII2023)*, Santa Fe, New Mexico, USA and Online.

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

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

Riccioni, I., R. Bongelli, et A. Zuczkowski (2021). Self-mention and uncertain communication in the British Medical Journal (1840-2007) : The decrease of subjectivity uncertainty markers. *Open Linguistics 7*(1), 739–759.

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

Walker, W., P. Harremoës, J. Rotmans, J. van der Sluijs, M. van Asselt, P. Janssen, et M. Krayer von Krauss (2003). Defining Uncertainty : A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment 4*(1), 5–17, doi: 10.1076/iaij.4.1.5.16466.

Summary

Uncertainty is an integral part of the scientific research process and is inherent in the construction of new knowledge. In this article, we examine the way in which uncertainty is expressed in scientific articles, and propose an annotation framework that takes into account the different dimensions of this notion. Scientific uncertainty is defined here as the expression of a lack of knowledge or a lack of precision in the information on an identified subject or concept. We propose a gold standard dataset composed of 1,839 sentences of manually annotated scientific articles from several disciplines. We also propose a linguistic knowledge-based approach for the automatic annotation of articles and for the detection and categorisation of scientific uncertainty. We compare the effectiveness of our approach in terms of Precision, Recall and F1 scores to the few-shot prompting methods performed via the Phi-3.5 and Llama 3 Large Language Models for the same annotation task. This comparative evaluation shows similar scores between the different approaches, with F1 scores up to 0.858 for our approach.