« De nos jours, ce sont les résultats qui comptent » : création et étude diachronique d'un corpus de revendications issues d'articles de TAL

Clémentine Bleuze[1] Fanny Ducel[2] Maxime Amblard[1] Karën Fort[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) Université Paris-Saclay, CNRS, LISN, F-91400 Orsay, France
clementine.bleuze@univ-lorraine.fr

RÉSUMÉ

Nous constituons un corpus de phrases issues de pré-tirages et d'articles de TAL, publiés en anglais entre 1952 et 2024, dont nous annotons manuellement un échantillon avec des catégories de revendications reflétant leur fonction rhétorique au sein des articles. Nous affinons un modèle SciBERT (Beltagy *et al.*, 2019) pour prédire les étiquettes restantes, que nous mettons, avec le corpus annoté, à la disposition de la communauté. Nous illustrons l'intérêt du corpus par des analyses exploratoires sur les caractéristiques des revendications relevées, ainsi qu'une étude diachronique de l'évolution de la structure des résumés ; ceci est mis en lien avec une réflexion sur la notion d'exagération scientifique. Nous observons une importance croissante des séquences de contexte précédant l'exposé des contributions, lequel est également de plus en plus suivi de séquences de résultats.

ABSTRACT

"Nowadays, the focus is on results" : creation and exploratory investigation of a corpus of claims from NLP articles.

We create a corpus of sentences from NLP papers and pre-prints published in English between 1952 and 2024, a sample of which we manually annotate with claim categories labels reflecting their rhetorical function. We fine-tune a SciBERT model (Beltagy *et al.*, 2019) to predict remaining labels, and make both the corpus and the model available to the community. We illustrate the interest of the corpus with exploratory analyses on the characteristics of identified claims, and a diachronic study of the evolution of abstracts' structures. We connect these analyses with a reflection about scientific overclaiming. We note a growing importance of context claims preceding contribution statements, which are themselves increasingly followed by sequences of result claims.

MOTS-CLÉS : zonage argumentatif, revendications, éthique, TAL pour le TAL.

KEYWORDS: argmentative zoning, claims, ethics, NLP4NLP.

1 Introduction

L'écriture de la recherche n'est ni neutre, ni purement objective. Si la relecture par les pairs permet généralement d'en corriger les imprécisions avant publication, la rédaction d'un article demeure un exercice éminemment rhétorique (Horton, 1995) dans lequel les chercheur·euses adoptent des stratégies pour émettre et défendre des revendications (*claims*) auprès du lectorat. L'étude de la

structure argumentative des articles permet alors une analyse des pratiques d'écriture et de publication d'une communauté donnée (ici alignée avec la démarche de « TAL pour le TAL » de Mariani *et al.* (2019a,b)), établissant un lien avec la question transverse de l'exagération scientifique. Dans un contexte où l'impératif de « publier ou périr » (*publish or perish*) a été qualifié de « mode de vie » du milieu académique (Rawat & Meena, 2014), il semble primordial de questionner nos propres recherches ainsi que la manière dont nous en rendons compte à nos pairs.

Dans cette optique, nous souhaitons dresser un panorama de la recherche écrite en TAL depuis plusieurs décennies, en nous intéressant à la nature des revendications rédigées par les auteur·ices. Dans un premier temps, nous envisageons l'idée qu'une quantification de la certitude des revendications de même nature peut servir de *proxy* pour détecter l'exagération, ce qui se révèle difficile en pratique (cf. Section 4.2). Dans un second temps, nous considérons que la nature des revendications exprimées ainsi que leur enchaînement au sein de structures stéréotypées constituent également des descripteurs intéressants de la manière dont on écrit la recherche en TAL (cf. Section 4.3).

Les contributions de ce travail consistent en : (i) la constitution d'un corpus de plus de 15 millions de phrases issues d'articles et pré-tirages de TAL en anglais publiés entre 1952 et 2024 ; (ii) l'annotation manuelle d'un échantillon du corpus en catégories de revendications selon une taxonomie créée par nos soins, inspirée des travaux de Zonage Argumentatif, et l'annotation automatisée de la partie restante par un modèle SciBERT (Beltagy *et al.*, 2019) affiné ; (iii) des analyses exploratoires sur les caractéristiques des différents types de revendication, leur niveau de certitude, ainsi qu'une étude diachronique de la structure des résumés d'articles ; (iv) l'ouverture d'une réflexion sur les pratiques d'écriture ainsi que l'exagération dans la recherche en TAL.

2 État de l'art

Enjeux rhétoriques de l'écriture scientifique : l'art et la manière de rédiger un titre (Nair & Gibbert, 2016), une introduction (Swales, 1981), une section de résultats (Thompson, 1993) et plus généralement un article scientifique entier (Patience *et al.*, 2015; Labaree, 2024) ont fait l'objet de multiples études et guides. Ces études décrivent (ou prescrivent) l'utilisation de procédés rhétoriques (*moves*) permettant d'articuler les revendications présentées et de justifier de l'intérêt de la recherche présentée [1]. « Faites attention en lisant cet article. Mon but est de persuader. », avertit ainsi Horton (1995) de manière quelque peu provocante avant d'évoquer l'emploi d'une voix active (ou passive), le choix et la position des adverbes et adjectifs, l'utilisation de la première (ou troisième) personne, ou encore l'auto-citation comme autant d'outils linguistiques permettant aux auteur·ices de moduler le « pouvoir persuasif profond » de leur démonstration. Martín-Martín (2008) étudie quant à lui l'atténuation (*hedging*) des revendications scientifiques comme une stratégie permettant aux auteur·ices de réduire le risque d'opposition, dans des articles en français et en espagnol.

Structure argumentative des articles et résumés : à un niveau plus superficiel, le Zonage Argumentatif (*Argumentative Zoning*) est « [l']analyse de la structure argumentative et rhétorique d'un article scientifique » par une classification phrase à phrase au sein de catégories mutuellement exclusives. Chez Teufel *et al.* (1999), on retrouve principalement les catégories de BUT, CONTEXTE, CONTRASTE, SUPPORT, REVENDICATION et AUTRE (catégories traduites en français). D'autres modèles existent cependant, dont Schrader *et al.* (2023) proposent un alignement (Figure 4) ayant

1. Par exemple, le modèle CARS (*Create A Research Space)* proposé par Swales (1990) propose de structurer une introduction autour de trois stratégies rhétoriques établissant la situation, le problème, puis la solution apportée.

nourri la réflexion autour de notre taxonomie (cf. Section 3.2). Le Zonage Argumentatif se décline également au niveau des résumés d'articles (*Sequential/Abstract Sentence Classification*), avec des approches automatisées utilisant des modèles de réseaux neuronaux hiérarchiques (Jin & Szolovits, 2018; Tokala *et al.*, 2023), des Champs Aléatoires Conditionnels (Yamada *et al.*, 2020) ou encore des modèles de langue pré-entraînés (Cohan *et al.*, 2019). Là encore, la granularité de la taxonomie utilisée peut varier. Il est important de noter que le champ médical, qui est fréquemment l'objet de ces travaux, se distingue du TAL par sa tradition d'employer des résumés structurés, généralement considérés comme plus lisibles et informatifs (Sharma & Harrison, 2006). Des études qualitatives concernant la structuration rhétorique de résumés d'articles ont également été proposées dans divers domaines (Sollaci & Pereira, 2004; Šauperl *et al.*, 2008).

3 Constitution d'un corpus annoté en types de revendications

3.1 Collecte des données de l'Anthologie ACL et ArXiv

L'Anthologie ACL [2] constitue une source incontournable d'articles de TAL publiés dans de multiples conférences (ACL, EMNLP, LREC, etc.). Nous proposons de l'enrichir avec des articles et pré-tirages publiés sur la plateforme ArXiv [3] qui, bien que non relue par les pairs, héberge également une abondante production scientifique, dont il est intéressant de se demander si les pratiques d'écriture diffèrent de celles de l'Anthologie ACL. Nous ré-utilisons le corpus ACL OCL (Rohatgi *et al.*, 2023) contenant les méta-données et le texte intégral au format XML de 71 286 articles publiés entre 1952 et 2022, et nous récupérons les méta-données d'articles [4] ArXiv (arXiv.org submitters, 2024) de la catégorie « linguistique computationnelle » (cs.CL) publiés entre 1992 et 2024, en écartant les doublons [5]. Nous téléchargeons ensuite les documents PDF associés, que nous convertissons au format XML avec l'outil GROBID (GRO, 2008 2024). Il est à noter qu'à ce stade, l'outil peine à effectuer la conversion de certains documents (notamment les plus anciens, de qualité visuelle réduite).

source	#articles identifiés	#articles parsés (XML)	#articles parsés (texte)	#phrases	#phrases/art.
Anthologie ACL	71 286	71 286	58 456	9 339 173	159,76
ArXiv	33 815	30 433	29 311	6 511 636	222,16
Total	105 101	101 719	87 767	15 850 809	180, 60

TABLE 1 – Détail du nombre d'articles et de phrases par source à chaque étape du pré-traitement.

Parmi les fichiers XML bien formés obtenus, nous excluons encore pour la phrase d'extraction ceux qui (i) correspondent à des posters ; (ii) ne sont pas rédigés en anglais. Finalement, pour chaque article restant, nous extrayons sa structure (les intitulés des sections et sous-sections) ainsi que son contenu textuel (hors figures, tableaux, références, etc.), ensuite segmenté en phrases à l'aide du modèle `en_core_web_sm` de la librairie `spaCy` [6]. La Table 1 récapitule des statistiques élémentaires sur le corpus ainsi constitué aux différentes étapes de pré-traitement.

2. https://aclanthology.org.
3. https://arxiv.org/.
4. Nous utilisons désormais *article* au sens large, incluant les articles relus par les pairs et les pré-tirages.
5. Le sous-corpus ArXiv comprend donc des articles publiés *uniquement* sur ArXiv, et non dans l'ACL Anthologie.
6. https://spacy.io/models/en#en_core_web_sm.

3.2 Annotation manuelle suivant une taxonomie de revendications

Dans un premier temps, nous utilisons un échantillon du corpus pour mettre au point une taxonomie de types de revendications non-mutuellement exclusifs, susceptibles de décrire les phrases des sections de résumé, introduction, résultats et conclusion des articles de TAL [7], inspirées par les catégories identifiées dans les travaux de Zonage Argumentatif. Cette taxonomie est progressivement affinée par ajout, suppression, ou fusion de catégories au cours de quatre phases d'annotation manuelle (voir Table 4) ayant impliqué six annotateur·ices travaillant sur la plateforme Doccano (Nakayama *et al.*, 2018). La version finale de cette taxonomie, stabilisée avec un coefficient α de Krippendorff (Krippendorff, 2011, 2013) de 0,81 entre les deux annotatrices principales, est présentée Table 2 (voir Table 8 pour des exemples de revendications annotées).

Catégorie	Caractéristiques
CONTEXTE	Éléments de contexte, d'état de l'art, d'explications théoriques préalables, etc.
CONTRIBUTION	Nature et caractéristiques des contributions de l'étude (un modèle, une ressource, une enquête, etc.), objectifs, méthodes, etc.
*PLAN	Pure description de la structure de l'article, de figures.
RÉSULTAT	Résultats (expérimentaux ou non), analyses, discussions, opinions des auteur·ices, hypothèses, etc.
IMPACT	Impact anticipé ou observé de l'étude sur des personnes, sur la communauté.
DIRECTIONS	Intentions de poursuites du travail, pistes de recherches futures.
LIMITATION	Limitations anticipées ou observées, défauts, imperfections du travail présenté.
NON-REVENDICATION	Phrases ne correspondant à aucune des catégories présentées ci-dessus.

TABLE 2 – Taxonomie des types de revendications applicables au TAL, identifiées lors de phases d'annotation manuelle. Nous notons que *PLAN n'est pas à proprement parler un type de revendication, cependant nous conservons cette catégorie notamment utile pour décrire la structuration des résumés et introductions.

Ces deux annotatrices ont ensuite procédé à l'annotation manuelle de 14 792 phrases issues de 158 articles [8], pour un total de 15 992 annotations (595 phrases ont plus d'une catégorie) et d'environ 27 h de travail. La distribution des annotations collectées par catégorie est présentée dans la Table 7. Bien que certaines catégories restent peu représentées (IMPACT ne compte que 154 phrases), ce premier sous-corpus constitue un jeu de données de qualité, relativement représentatif de notre corpus de littérature en TAL.

3.3 Entraînement de modèles pour la classification des revendications

L'échantillon manuellement annoté est ensuite utilisé comme corpus d'entraînement pour la classification automatique des revendications, vue comme un problème de classification multi-étiquettes à huit classes au niveau de la phrase. Nous comparons la performance de modèles « traditionnels » de l'apprentissage automatique (par opposition aux réseaux de neurones profonds), ainsi que plusieurs

7. Nous considérons que les autres sections d'un article sont moins susceptibles de contenir des revendications majeures.

8. Ces articles sont sélectionnés de manière à représenter à la fois l'ACL Anthologie (52,5 %) et ArXiv (47,5 %) ; ainsi que les différentes périodes de publication : <1994 (15.2 %), 1994-2004 (29.7 %), 2004-2014 (27,2 %), >2014 (27,8 %). En réalité, le nombre d'articles croit quasi-exponentiellement avec l'année de publication, toutefois cette distribution forcée nous permet un rééquilibrage.

grands modèles pré-entraînés basés sur BERT (Devlin *et al.*, 2019), communément employés pour ce type de tâche. Coan *et al.* (2021) observent par exemple qu'un modèle de Régression Logistique associé à RoBERTa obtient les meilleures performances (macro F1-score de 0,79) dans un problème de classification avec des classes déséquilibrées. Ces modèles ont également l'avantage d'être relativement légers à implémenter, ce qui limite la puissance computationnelle nécessaire et, par conséquent, l'impact environnemental associé.

Modèles « traditionnels » : nous utilisons les implémentations `scikit-learn` [9] pour la Régression Logistique (LR) et les Machines à Vecteurs de Support (SVM) [10]. Nous comparons une vectorisation par sacs de mots (*CountVectorizer* ou CV) et par TF-IDF, ainsi que différents formats pour l'entrée : la phrase cible, ou la phrase précédée par l'intitulé de sa section dans l'article.

Modèles basés sur BERT : nous utilisons des modèles RoBERTa (Liu *et al.*, 2019), DeBERTa (He *et al.*, 2021) et SciBERT (Beltagy *et al.*, 2019) disponibles sur HuggingFace [11], dont nous effectuons l'affinement en 15 époques avec une stratégie d'arrêt anticipé pour éviter le sur-apprentissage. En entrée, nous fournissons la phrase cible précédée de l'intitulé de sa section (i) seule ; (ii) avec la phrase précédente ainsi que la phrase suivante ; (iii) avec les deux phrases précédentes.

modèle	CONT.	CONTR.	PLAN	RÉS.	IMP.	DIR.	LIM.	NR	F1 moyen
SVM (CV, section)	0,70	0,57	0,67	0,64	0,11	0,47	0,18	0,81	0,68
SciBERT (2 préc.)	**0,93**	**0,87**	**0,88**	**0,86**	**0,52**	**0,82**	**0,51**	**0,95**	**0,89**

TABLE 3 – F1-scores de classification des différentes classes pour les deux meilleurs modèles de chaque famille : SVM avec vectorisation en sacs de mots et utilisation de l'intitulé de section, et SciBERT avec contexte intégrant les deux phrases précédant la phrase cible.

Les modèles sont comparés en fonction de leur F1-score pondéré (voir résultats complets Tables 5 et 6, et les meilleurs pour chaque famille de modèles Table 3). Le modèle SciBERT obtient le meilleur F1-score moyen pondéré (0,89), avec une performance optimale pour la classe CONTRIBUTION (0,93) et le plus de difficultés pour les classes IMPACT et LIMITATION (0,52 et 0,51), qui représentent une faible part du corpus d'apprentissage. Nous utilisons ce modèle pour inférer les classes des phrases restantes du corpus, bien que ce dernier point souligne que la fiabilité de ces « pseudo-étiquettes » (*silver labels*) ne soit pas complètement assurée pour les classes moins représentées. La distribution obtenue est comparée avec celle issue de l'annotation manuelle en Table 7. Nous fournissons l'intégralité du corpus (méta-données des articles, fichiers XML, phrases issues des articles extraits, annotations manuelles et automatiques) en libre accès [12], ainsi que le modèle SciBERT affiné ayant effectué les annotations automatiques [13].

Vérification manuelle (partielle) de la qualité des prédictions : pour nous assurer de la fiabilité des étiquettes prédites par le modèle, nous tirons aléatoirement 100 articles dont les étiquettes des phrases de résumés [14] sont comparées à des annotations effectuées manuellement par une annotatrice. Sur les 604 phrases annotées, nous relevons 11 phrases issues d'erreurs de *parsing* (notes de bas de page,

9. https://scikit-learn.org/stable/.
10. Pour ce dernier modèle, une optimisation par quadrillage fixe un noyau sigmoïde et $C = 5$.
11. https://huggingface.co/FacebookAI/roberta-base ; https://huggingface.co/microsoft/deberta-v3-base ; https://huggingface.co/allenai/scibert_scivocab_uncased.
12. https://huggingface.co/datasets/ClementineBleuze/CNP ; https://github.com/ClementineBleuze/claims-in-NLP
13. https://huggingface.co/ClementineBleuze/scibert_claim-classification
14. Il nous est impossible de vérifier l'ensemble des prédictions ; ici nous focalisons notre attention sur les résumés en anticipation de l'analyse présentée Section 4.3.

début d'introduction ne devant en théorie pas être retenu dans le texte de résumé), exclues pour ce qui suit. Les annotations sont strictement identiques dans 87,35 % des cas, pour un α de Krippendorff de 0,83, ce qui constitue un accord solide [15]. Les désaccords portent sur une séquence de plusieurs revendications de CONTEXTE mal identifiées comme CONTRIBUTION dans un même article, et sur la multi-étiquette CONTRIBUTION+RÉSULTAT alternativement annotée plutôt CONTRIBUTION ou RÉSULTAT par l'autre partie. Dans l'ensemble, ces observations confirment la difficulté de la catégorisation dans certaines situations ambiguës, tout en permettant d'accorder un crédit raisonnable aux étiquettes prédites.

4 Analyses sur le corpus

4.1 Analyses exploratoires sur les différents types de revendications

En excluant les phrases identifiées comme NON REVENDICATION, notre corpus contient un total de 5 592 131 revendications, dont 302 213 comptent plus d'une étiquette (voir distribution Table 7). Afin d'étudier plus en détail l'ensemble de ces phrases, nous en relevons le vocabulaire, les entités nommées et catégories grammaticales à l'aide du modèle `en_core_web_sm` [16], ainsi que des indicateurs comme la longueur des phrases et leur nombre de caractères.

Co-occurrences de catégories : nous observons que de nombreuses phrases comptent plus d'une étiquette (voir exemples Table 9). Ainsi, les couples de revendications qui co-occurrent le plus fréquemment sont LIMITATION+RÉSULTAT (39 % des occurrences de LIMITATION), IMPACT+DIRECTIONS (14 % des occurrences d'IMPACT) et PLAN+CONTRIBUTION (13 % des occurrences de PLAN) [17]. Les deux premiers couples concernent des catégories parfois sujettes à ambiguïté lors des phases d'annotations manuelles, avec la difficulté de distinguer un résultat négatif d'une « pure limitation » ou encore une revendication d'impact avec une prédiction ou anticipation formulée au futur.

Spécificités linguistiques : si certains termes ou bigrammes identifient assez nettement une catégorie (par ex. « travaux futurs » pour DIRECTIONS ou « section » pour PLAN), les vocabulaires des autres types de revendications se recoupent assez largement autour de termes comme « modèle » et « utiliser » (présents dans le top-10 de toutes les catégories), ce qui peut expliquer les performances modérées d'une approche basée sur les sacs de mots (cf. résultats Table 3). Nous notons que les revendications les plus longues sont celles d'IMPACT (env. 169 caractères, 29 tokens/phrase) et les plus courtes les annonces de PLAN (env. 100 caractères, 18 tokens/phrase). Les entités nommées de type CARDINAL sont davantage présentes dans les revendications de CONTEXTE, CONTRIBUTION et RÉSULTAT, tandis que celles de type DATE, ENTITÉ GÉOPOLITIQUE et PERSONNE sont quasi-exclusives aux revendications de CONTEXTE. Enfin, les négations sont environ trois fois plus présentes dans les LIMITATIONS que dans les RÉSULTATS. Ceci met en lumière des indices linguistiques permettant de discriminer entre certains types de revendications.

15. Cet accord est d'ailleurs supérieur à l'accord inter-annotatrices à l'issue des phases d'annotation manuelle ayant permis de constituer la taxonomie, qui était de 0,81. Il semblerait que sur une tâche complexe d'annotation multi-classes et multi-étiquettes, on approche d'un « plafond » d'accord difficilement dépassable.
16. https://spacy.io/models/en#en_core_web_sm.
17. Noter que les co-occurrences sont orientées.

4.2 Niveau de certitude : un indicateur limité

Suivant une hypothèse initiale de notre travail [18], nous utilisons des modèles proposés par Pei & Jurgens (2021) pour mesurer la certitude de l'ensemble des phrases du corpus. Également basés sur SciBERT, ceux-ci fournissent, pour une phrase donnée : un score global de certitude (*sentence-level certainty*) compris entre 1 et 6 (6 étant la certitude maximale), ainsi que des indicateurs de certitude binaires reliés à des aspects (*aspect-level certainty*). Par exemple, l'aspect de QUANTITÉ peut être certain dans « une précision de 98 % », mais incertain dans « une précision d'*environ* 98 % ».

Cependant, en inspectant les prédictions, nous constatons que la distribution du score de certitude global a une amplitude très réduite ($Q_1 \approx 4{,}75$; $Q_3 \approx 5{,}00$), et que les phrases aux scores les plus extrêmes nous paraissent peu convaincantes (cf. Table 10). En ce qui concerne la certitude basée sur les aspects (QUANTITÉ, DEGRÉ, PROBABILITÉ, SUGGESTION, CADRAGE et CONDITION), nous observons des tendances globales crédibles (par ex. un aspect de PROBABILITÉ davantage "Incertain" dans les revendications de LIMITATION que de RÉSULTATS) mais des difficultés de lisibilité au niveau des phrases elles-mêmes. Nous regrettons par ailleurs que les prédictions ne s'accompagnent pas d'indications sur les parties de la phrases ayant motivé les résultats, ce qui les rend très peu lisibles [19]. Au vu de l'ensemble de ces observations, nous renonçons à exploiter ces annotations comme indicateurs de l'exagération, et décidons d'orienter nos analyses vers l'étude des structures au sein desquelles les revendications sont ordonnées.

4.3 Étude diachronique de la structure des résumés en TAL

En guise d'exemple d'utilisation possible du corpus pour une analyse qualitative des pratiques d'écriture en TAL, nous proposons d'étudier la structure des résumés du corpus au cours du temps, c'est-à-dire les enchaînements de séquences (revendications consécutives partageant la même étiquette) les plus fréquents (voir par ex. Figure 1). Nous motivons notre intérêt envers les résumés en ce qu'ils constituent une sorte de « vitrine » des articles, mais ceci pourrait naturellement être étendu à d'autres sections typiques telles que les introductions ou conclusions.

Les résultats s'imposent dans les résumés : un premier constat est que la diversité des types de revendications ne se retrouve pas dans les résumés : ils sont seulement 8,98 %, 3,43 % et 2,57 % à compter des revendications d'IMPACT, de LIMITATION et de DIRECTIONS. Pourtant, à l'échelle de l'article entier, ces mêmes catégories sont présentes dans plus de 45 %, 70% et 78 % des cas. Ceci confirme le rôle particulier des résumés pour mettre en avant certains types de revendications-clés, au détriment par exemple des limitations identifiées par les chercheur·euses dans leur propre travail, davantage détaillées dans les sections de résultats, d'analyse, ou de discussion. Nous observons d'autre part un changement de paradigme assez important concernant la présence de revendications de RÉSULTATS dans les résumés : 89,95 % des articles publiés en 2024 en contiennent, contre seulement

18. Un de nos premiers objectifs était d'étudier les revendications issues d'articles de recherche, dans le but de détecter d'éventuelles exagérations. Une piste évoquée était ainsi de mesurer l'exagération comme une simple différence numérique de score de certitude entre des paires d'énoncés censés exprimer des contenus similaires (par exemple, des résultats scientifiques tels qu'annoncés en introduction puis repris dans la conclusion d'un même article). Un procédé semblable de différence de scores de certitude est utilisé par Patro & Baruah (2021).

19. À titre d'exemple, une phrase telle que : « La meilleure performance atteinte était de 29 % en F-score, tandis que beaucoup d'équipes ont obtenu un score sous les 10 % » aurait ainsi des aspects de NOMBRE, DEGRÉ et PROBABILITÉ "Certains" (les autres aspects n'ont pas été identifiés dans la phrase), ce qu'il nous est difficile de justifier avec confiance de manière *post-hoc*.

42,86 % en 1978. La tendance est nette et continue au cours du temps (Figure 2), et pourrait témoigner d'un passage en TAL vers un « paradigme d'évaluation » reposant sur un recours systématisé aux *benchmarks* pour évaluer les modèles.

Contexte	Les relations discursives relient des unités linguistiques plus petites en textes cohérents.
Contexte	L'identification automatique des relations discursives est difficile, car elle nécessite de comprendre la sémantique des arguments liés entre eux.
Contexte	Un défi plus subtil réside dans le fait qu'il ne suffit pas de représenter le sens de chaque argument d'une relation discursive, car celle-ci peut dépendre de liens entre des composants de niveau inférieur, tels que les mentions d'entités.
Contribution	Notre solution calcule des représentations de sens distribuées pour chaque argument discursif par composition ascendante dans l'arbre d'analyse syntaxique.
Contribution	Nous effectuons également une passe compositionnelle descendante pour capturer le sens des mentions d'entités coréférentes.
Résultat	Les relations discursives implicites sont ensuite prédites à partir de ces deux représentations, ce qui permet d'obtenir des améliorations substantielles sur le *Penn Discourse Treebank*.

FIGURE 1 – Exemple d'un résumé dont la structure en CONTEXTE-CONTRIBUTION-RÉSULTAT est typique des articles du corpus publiés entre 2010 et 2024. Résumé de Ji & Eisenstein (2015), traduit en français.

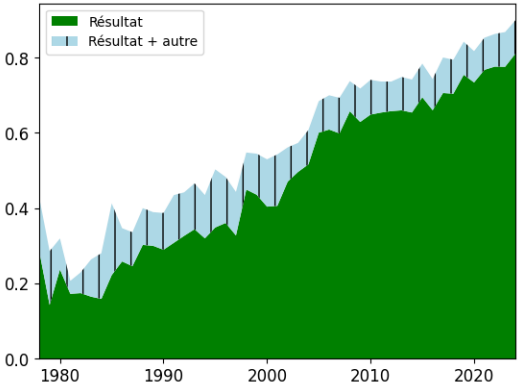

FIGURE 2 – Part des articles du corpus comprenant au moins une séquence RÉSULTAT dans leur résumé en fonction de l'année de publication. Nous considérons également les multi-étiquettes comprenant RÉSULTAT. Par souci de lisibilité, nous commençons en 1978, année à partir de laquelle le corpus compte plus de 10 articles par an.

Évolution de la structure des résumés : nous proposons en Figure 3 une visualisation de la construction des résumés au cours de trois périodes consécutives (avant 1995, 1995-2009, et 2010-2024) à l'aide de diagrammes de Sankey. Nous observons premièrement une normalisation progressive des pratiques d'écriture, avec un nombre d'options de plus en plus limité au cours du temps pour les séquences arrivant en position 1 à 3, ainsi qu'une inversion de la tendance CONTRIBUTION/CONTEXTE pour la première position. Nous retrouvons l'augmentation de la place prise par les RÉSULTATS discutée dans le point précédent, ainsi qu'un renforcement des alternances CONTRIBUTION-RÉSULTAT

au cours du résumé. Enfin, les structures s'allongent : si la structure majoritaire avant 1995 ne compte qu'une séquence de CONTRIBUTION, elle passe entre 1995 et 2009 à CONTRIBUTION-RÉSULTAT, puis entre 2010 et 2024 à CONTEXTE-CONTRIBUTION-RÉSULTAT. Des résumés typiques de chaque période sont présentés en Figures 1, 5 et 6. Notons qu'il serait intéressant de compléter ces premières observations structurelles par des considérations plus fines portant par exemple, à l'intérieur des séquences, sur la polarité du vocabulaire employé, à la manière de Vinkers *et al.* (2015).

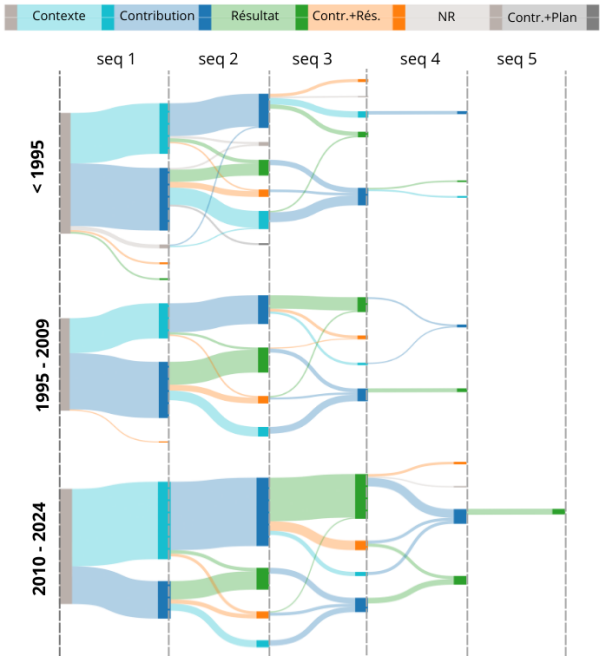

FIGURE 3 – Diagrammes de Sankey permettant de visualiser la distribution des structures de résumés du corpus à plusieurs périodes. Les diagrammes se lisent de gauche à droite, et la largeur des flux est proportionnelle à la part des séquences de revendications correspondantes à chaque étape de construction du résumé. Par exemple, entre 2010 et 2024, les résumés commencent par une séquence de CONTEXTE dans 67,4 % des cas, et de CONTRIBUTIONS dans les 32,6 % restants. Ceux commençant par du CONTEXTE enchaînent majoritairement avec une séquence de CONTRIBUTIONS, tandis que ceux ayant commencé par les CONTRIBUTIONS sont principalement suivis de RÉSULTATS, etc. Les interruptions de flux signalent la fin des résumés (dont la plupart comportent au plus trois séquences différentes consécutives, même si certains en ont jusqu'à cinq). Par souci de lisibilité, les flux représentant moins de 1 % des effectifs d'un noeud sortant n'ont pas été représentés.

5 Discussion : pratiques d'écriture et exagération en TAL

Dans ce qui suit, nous revenons sur plusieurs questionnements transverses à l'étude des revendications, notamment en ce qui concerne l'exagération scientifique. Nous identifions également des pistes qui nous semblent intéressantes à approfondir dans le cadre de travaux futurs.

L'exagération ne porte pas exclusivement sur les résultats : notre étude a été initialement motivée par la volonté de modéliser l'exagération scientifique dans le contexte des articles de TAL. À la différence des résultats présentés dans les résumés d'articles médicaux, qui peuvent être « embellis [20] » (Koroleva & Paroubek, 2018; Koroleva, 2020), les revendications portées en TAL figurent dans des résumés non-structurés ne comportant pas d'entités caractéristiques propres au domaine, qui seraient le pendant des résultats primaires, secondaires, population cible, etc. ; quelles seraient alors, en TAL, les « revendications valant d'être vérifiées » ? C'est en tentant de répondre à cette question que nous avons entamé l'élaboration de notre taxonomie (cf. Section 3.2). Intuitivement, l'« exagération » évoquerait des cas d'enthousiasme excessif au sujet de RÉSULTATS ; par exemple, interpréter une augmentation de score BLEU de 0,5 % (ce qui est très faible) d'une méthode comme témoignant d'une performance « significativement meilleure ». Il nous semble cependant que l'exagération peut également s'insérer dans des CONTRIBUTIONS abusives (par ex. qualifier un modèle de langue de « réellement multilingue » sans fournir d'évaluation sur certaines langues pourtant prétendument prises en charge, ce qui est par exemple regretté par Jouitteau & Grobol (2024) pour le cas du Breton) ou des revendications d'IMPACT (par ex. suggérant qu'une étude de portée limitée participe de l'avènement d'une « IA générale » [21]). Ainsi, chacune des catégories retenues dans notre taxonomie nous semble pouvoir relever de formes spécifiques d'exagération. La question se pose bien sûr de savoir s'il est possible de qualifier consensuellement une revendication donnée d'« exagérée », et nous ne prétendons pas disposer d'une expertise définitive en la matière. Toutefois, nous espérons que notre travail pourra apporter sa contribution à l'élaboration d'une réflexion sur la manière dont nous, chercheurs et chercheuses en TAL, abordons, écrivons, et promouvons notre recherche.

Une écriture différente dans l'Anthologie ACL et ArXiv ? : jusqu'ici, nous n'avons pas pris en compte l'origine des articles étudiés. Ceux issus de l'Anthologie ACL et d'ArXiv ne recouvrent pas les mêmes périodes de publication : ils s'étendent respectivement de 1952 à 2022 et de 1994 à 2024, cependant les distributions ne sont réellement semblables qu'entre 2010 et 2022 (cf. Figure 7). Une comparaison synchronique des structures des résumés sur cette période montre en Figure 8 de grandes similitudes, si ce n'est pour la part occupée par les résumés commençant par une séquence CONTEXTE, encore davantage préférée dans le corpus ArXiv à ceux commençant par une séquence de CONTRIBUTIONS. Il semblerait donc qu'une certaine « narration contextualisée » soit plus caractéristique de la plateforme ArXiv, indépendamment des évolutions diachroniques relevées précédemment. À l'échelle de l'article entier, nous notons également qu'il est significativement plus fréquent d'observer des revendications d'IMPACT sur ArXiv (55,18 % des articles) que dans l'ACL Anthologie (45,96 %), et à l'inverse que les phrases de PLAN y sont moins fréquentes (42,63 % contre 49,14 %). Ceci pourrait provenir d'une moins grande formalité dans les pré-tirages ou d'une tendance à davantage mettre en avant des résultats « impactants » ; mais ce dernier point est à nuancer par la moins bonne performance du modèle pour identifier les revendications d'IMPACT. La question du style propre à chacune de ces sources de littérature mériterait selon nous d'être approfondie par des travaux supplémentaires.

Une écriture qui se normalise : il est intéressant de constater que la structure CONTEXTE-CONTRIBUTION-RÉSULTAT désormais prédominante dans les résumés de TAL semble s'aligner avec la structure « Introduction, Méthodes, Résultats et Discussion » (IMRAD) généralement observée à

20. L'« embellissement » (*spin*) désigne le fait de « présenter des résultats de recherche d'une manière plus favorable (ou, rarement, plus défavorable) que ce que suggèrent les preuves obtenues » (Koroleva, 2020). Selon Boutron *et al.* (2014), ceci peut faire surestimer à des cliniciens l'efficacité d'un traitement expérimental.

21. Voir par exemple les revendications d'impact très optimistes de Rosenberg *et al.* (2023) au sujet d'une technologie ayant fait ses preuves sur une unique tâche : celle d'estimer le nombre de friandises contenues dans une jarre, à partir d'une photographie.

l'échelle de l'article entier, et majoritairement adoptée dans les sciences de la santé depuis les années 1970 (Sollaci & Pereira, 2004). En s'alignant sur ce standard issu d'une autre discipline scientifique, la littérature en TAL propose des résumés quasi-structurés, bien que la rédaction de cette partie demeure libre dans les faits. On peut également associer ce phénomène à la visibilité croissante du domaine : alors que des milliers d'articles sont désormais recensés chaque année dans l'Anthologie ACL, les publications plus anciennes visaient un public restreint, composé de spécialistes. Il semble dans ce cas naturel qu'un exposé des CONTRIBUTIONS d'un travail puisse se passer d'une séquence de CONTEXTE introductive ; de nos jours, l'explosion du nombre de disciplines et sous-disciplines liées au TAL rendrait un tel résumé moins accessible. Par ailleurs, la nette augmentation de la part des RÉSULTATS au sein des résumés pourrait être une conséquence observable de l'évolution du TAL depuis les années 1990 vers un paradigme d'évaluation faisant la part belle aux *benchmarks*, ce qui peut notamment se traduire par des incitations à publier des résultats toujours meilleurs que l'état de l'art à un moment donné. Il est intéressant de noter que cette forme de course à la performance, qui conduit à une homogénéisation des publications, risque à terme de détourner les chercheur·euses de travaux potentiellement plus impactants pour la communauté scientifique. Selon Church & Kordoni (2022), cette pratique risquerait également de décourager les approches disciplinaires tout en favorisant l'apparition de revendications exagérées autour de la performance prétendument « surhumaine » des modèles les plus performants.

6 Conclusion

Nous constituons et mettons librement à disposition de la communauté un corpus de plus de 15 millions de phrases issues d'articles et de pré-tirages publiés dans le domaine du TAL (en anglais) entre 1952 et 2024, avec une annotation en catégories de revendications. Un échantillon de ces annotations est effectué manuellement, le reste est prédit par un modèle SciBERT affiné, également mis à disposition. Nous illustrons l'intérêt de ce corpus en proposant plusieurs analyses exploratoires, dont une étude diachronique de la structure des résumés. Nous montrons qu'aux simples exposés de contributions caractéristiques des articles antérieurs à 1995 se sont progressivement greffées des séquences de résultats (1995-2009), puis des séquences de contexte en première position (depuis 2010). Nous accompagnons ces observations d'une réflexion sur l'évolution des pratiques d'écriture en TAL, ainsi que sur le phénomène d'exagération scientifique ayant initialement motivé ce travail. Dans la poursuite de cette réflexion, des travaux futurs pourraient par exemple se pencher sur les structures des conclusions, ou approfondir l'examen des différences d'écriture entre articles publiés dans l'Anthologie ACL et dans la plateforme de pré-tirages ArXiv.

Limites - Des imperfections ont pu s'insérer dans les données du corpus, à chaque étape de traitement : documents PDF manquants, difficultés de *parsing* en XML (en particulier pour les articles les plus anciens), erreurs de segmentation, d'annotation automatique, etc. L'annotation manuelle a également révélé qu'un « plafond d'accord » semblait exister pour la tâche complexe de catégorisation de revendications avec plusieurs étiquettes possibles. Enfin, nous n'avons considéré que des articles rédigés en anglais, ce qui ne recouvre pas l'intégralité de la littérature disponible en TAL.

Remerciements

Nous souhaitons remercier le programme ORION de l'Université de Lorraine [22] (ANR-20-SFRI-0009), ainsi que le projet ANR InExtenso [23] (ANR-23-IAS1-0004) qui ont contribué au financement de ce travail. Nous avons également bénéficié d'un accès aux ressources de la plateforme de calcul Grid5000 [24]. Merci enfin aux annotateur·ices supplémentaires ayant apporté leur aide aux auteur·ices lors de la phase de constitution de notre taxonomie : Amandine Decker (CLASP, Université de Göteborg, et LORIA, Université de Lorraine) et Valentin Richard (LORIA, Université de Lorraine, et ILLC, Université d'Amsterdam).

Références

(2008–2024). Grobid. https://github.com/kermitt2/grobid.

arXiv.org submitters (2024). arxiv dataset. DOI : 10.34740/KAGGLE/DSV/7548853.

Beltagy I., Lo K. & Cohan A. (2019). SciBERT : A pretrained language model for scientific text. In K. Inui, J. Jiang, V. Ng & X. Wan, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3615–3620, Hong Kong, Chine : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1371.

Bhat I., Bhat R. A., Shrivastava M. & Sharma D. (2018). Universal Dependency parsing for Hindi-English code-switching. In M. Walker, H. Ji & A. Stent, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 987–998, La Nouvelle-Orléans, Louisiane : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1090.

Boutron I., Altman D. G., Hopewell S., Vera-Badillo F., Tannock I. & Ravaud P. (2014). Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer : the spiin randomized controlled trial. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, **32**(36), 4120–4126. DOI : 10.1200/JCO.2014.56.7503.

Church K. W. & Kordoni V. (2022). Emerging trends : Sota-chasing. *Natural Language Engineering*, **28**(2), 249–269. DOI : 10.1017/S1351324922000043.

Coan T. G., Boussalis C., Cook J. & Nanko M. O. (2021). Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, **11**(1), 22320. DOI : 10.1038/s41598-021-01714-4.

Cohan A., Beltagy I., King D., Dalvi B. & Weld D. (2019). Pretrained language models for sequential sentence classification. In K. Inui, J. Jiang, V. Ng & X. Wan, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3693–3699, Hong Kong, Chine : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1383.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

22. https://www.univ-lorraine.fr/lue/orion/
23. https://anr-inextenso.loria.fr/
24. https://www.grid5000.fr/w/Grid5000:Home.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DINARELLI M. & GROBOL L. (2019). Seq2Biseq : Bidirectional Output-wise Recurrent Neural Networks for Sequence Modelling. In *CICLing 2019 - 20th International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France. HAL : hal-02085093.

ERYIĞIT G., NIVRE J. & OFLAZER K. (2008). Dependency parsing of Turkish. *Computational Linguistics*, **34**(3), 357–389. DOI : 10.1162/coli.2008.07-017-R1-06-83.

FAN L., KRISHNAN D., ISOLA P., KATABI D. & TIAN Y. (2023). Improving clip training with language rewrites. In A. OH, T. NAUMANN, A. GLOBERSON, K. SAENKO, M. HARDT & S. LEVINE, Éds., *Advances in Neural Information Processing Systems*, volume 36, p. 35544–35575, La Nouvelle-Orléans, Louisiane : Curran Associates, Inc.

GAO S., TAKANOBU R., BOSSELUT A. & HUANG M. (2022). End-to-end task-oriented dialog modeling with semi-structured knowledge management. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **30**, 2173–2187. DOI : 10.1109/TASLP.2022.3153255.

GLASS J. & SENEFF S. (2003). Flexible and personalizable mixed-initiative dialogue systems. In *Proceedings of the HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing*, p. 19–21, Edmonton, Canada.

GUREVYCH I., MALAKA R., PORZEL R. & ZORN H.-P. (2003). Semantic coherence scoring using an ontology. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, p. 88–95, Edmonton, Canada.

HE P., LIU X., GAO J. & CHEN W. (2021). Deberta : Decoding-enhanced bert with disentangled attention. arXiv :2006.03654 [cs], DOI : 10.48550/arXiv.2006.03654.

HEADDEN III W. P., MCCLOSKY D. & CHARNIAK E. (2008). Evaluating unsupervised part-of-speech tagging for grammar induction. In D. SCOTT & H. USZKOREIT, Éds., *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, p. 329–336, Manchester, Royaume-Uni : Coling 2008 Organizing Committee.

HORTON R. (1995). The rhetoric of research. *BMJ*, **310**(6985), 985–987. DOI : 10.1136/bmj.310.6985.985.

JAMISON E. & GUREVYCH I. (2013). Headerless, quoteless, but not hopeless ? using pairwise email classification to disentangle email threads. In R. MITKOV, G. ANGELOVA & K. BONTCHEVA, Éds., *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, p. 327–335, Hissar, Bulgarie : INCOMA Ltd. Shoumen, Bulgarie.

JI Y. & EISENSTEIN J. (2015). One vector is not enough : Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, **3**, 329–344. DOI : 10.1162/tacl_a_00142.

JIN D. & SZOLOVITS P. (2018). Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3100–3109, Bruxelles, Belgique : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1349.

JOUITTEAU M. & GROBOL L. (2024). Petits oublis, grands effets : le silençage des communauté linguistiques minorisées dans le tal et ses conséquences. In *Journée d'étude Journée Ethique et TAL 2024*, Nancy, France. HAL : hal-04551943.

KILBURY J., BONTCHEVA K. & SAMIH Y. (2011). FTrace : A tool for finite-state morphology. In A. MALETTI & M. CONSTANT, Éds., *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, p. 88–92, Blois, France : Association for Computational Linguistics.

KOROLEVA A. (2020). *Assisted authoring for avoiding inadequate claims in scientific reporting*. phdthesis, Université Paris-Saclay ; Université d'Amsterdam.

KOROLEVA A. & PAROUBEK P. (2018). Annotating spin in biomedical scientific publications : the case of random controlled trials (RCTs). In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon : European Language Resources Association (ELRA).

KRIPPENDORFF K. (2011). Computing krippendorff's alpha-reliability.

KRIPPENDORFF K. (2013). *Content Analysis : An Introduction to Its Methodology*. Thousand Oaks, CA : Sage, 3rd edition édition.

LABAREE R. V. (2024). Research guides : Organizing your social sciences research paper. url : https ://libguides.usc.edu/writingguide/.

LEE N., LI B. Z., WANG S., YIH W.-T., MA H. & KHABSA M. (2020). Language models as fact checkers ? In C. CHRISTODOULOPOULOS, J. THORNE, A. VLACHOS, O. COCARASCU & A. MITTAL, Éds., *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, p. 36–41, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2020.fever-1.5.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. arXiv :1907.11692 [cs], DOI : 10.48550/arXiv.1907.11692.

LUKEŠ D., KOPŘIVOVÁ M., KOMRSKOVÁ Z. & POUKAROVÁ P. (2018). Pronunciation variants and ASR of colloquial speech : A case study on Czech. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon : European Language Resources Association (ELRA).

MARIANI J., FRANCOPOULO G. & PAROUBEK P. (2019a). The nlp4nlp corpus (i) : 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, **3**. DOI : 10.3389/frma.2018.00036.

MARIANI J., FRANCOPOULO G., PAROUBEK P. & VERNIER F. (2019b). The nlp4nlp corpus (ii) : 50 years of research in speech and language processing. *Frontiers in Research Metrics and Analytics*, **3**. DOI : 10.3389/frma.2018.00037.

MARTÍN-MARTÍN P. (2008). The mitigation of scientific claims in research papers : A comparative study. *IJES, International journal of english studies, ISSN 1578-7044, Vol. 8, Nº. 2, 2008 (Ejemplar dedicado a : Academic Writing : The Role of Different Rhetorical Conventions), pags. 133-152*, **8**. DOI : 10.6018/ijes.8.2.49201.

MERZ M. & SCRIVNER O. (2022). Discourse on ASR measurement : Introducing the ARPOCA assessment tool. In S. LOUVAN, A. MADOTTO & B. MADUREIRA, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 366–372, Dublin, Irlande : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-srw.28.

NAIR L. B. & GIBBERT M. (2016). What makes a 'good' title and (how) does it matter for citations ? a review and general model of article title attributes in management science. *Scientometrics*, **107**(3), 1331–1359. DOI : 10.1007/s11192-016-1937-y.

NAKAYAMA H., KUBO T., KAMURA J., TANIGUCHI Y. & LIANG X. (2018). doccano : Text annotation tool for human. Software available from https ://github.com/doccano/doccano.

OCH F. J. & NEY H. (2001). What can machine translation learn from speech recognition ? In S. KRAUWER, Éd., *Workshop on MT2010 : Towards a Road Map for MT*, Saint-Jacques-de-Compostelle, Espagne.

PATIENCE G. S., BOFFITO D. C. & PATIENCE P. A. (2015). How do you write and present research well ? *The Canadian Journal of Chemical Engineering*, **93**(10), 1693–1696. DOI : 10.1002/cjce.22261.

PATRO J. & BARUAH S. (2021). A simple three-step approach for the automatic detection of exaggerated statements in health science news. In P. MERLO, J. TIEDEMANN & R. TSARFATY, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 3293–3305, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.289.

PEI J. & JURGENS D. (2021). Measuring sentence-level and aspect-level (un)certainty in science communications. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 9959–10011, En ligne et Punta Cana, République Dominicaine : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.784.

PIERACCINI R., LEVIN E. & LEE C.-H. (1991). Stochastic representation of conceptual structure in the ATIS task. In *Speech and Natural Language : Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, Pacific Grove, Californie.

RANASINGHE T., ORASAN C. & MITKOV R. (2020). Intelligent translation memory matching and retrieval with sentence encoders. In A. MARTINS, H. MONIZ, S. FUMEGA, B. MARTINS, F. BATISTA, L. COHEUR, C. PARRA, I. TRANCOSO, M. TURCHI, A. BISAZZA, J. MOORKENS, A. GUERBEROF, M. NURMINEN, L. MARG & M. L. FORCADA, Éds., *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 175–184, Lisbonne, Portugal : European Association for Machine Translation.

RAWAT S. & MEENA S. (2014). Publish or perish : Where are we heading ? *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, **19**(2), 87–89.

ROHATGI S., QIN Y., AW B., UNNITHAN N. & KAN M.-Y. (2023). The ACL OCL corpus : Advancing open science in computational linguistics. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 10348–10361, Singapour : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.640.

ROSENBERG L., WILLCOX G. & SCHUMANN H. (2023). Towards collective superintelligence, a pilot study. In *2023 International Conference on Human-Centered Cognitive Systems (HCCS)*, p. 1–6, Cardiff, Royaume-Uni : IEEE. DOI : 10.1109/HCCS59561.2023.10452485.

SCHRADER T., BÜRKLE T., HENNING S., TAN S., FINCO M., GRÜNEWALD S., INDRIKOVA M., HILDEBRAND F. & FRIEDRICH A. (2023). Mulms-az : An argumentative zoning dataset for the materials science domain. In M. STRUBE, C. BRAUD, C. HARDMEIER, J. J. LI, S. LOAICIGA & A. ZELDES, Éds., *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, p. 1–15, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.codi-1.1.

SHARMA S. & HARRISON J. E. (2006). Structured abstracts : do they improve the quality of information in abstracts ? *American Journal of Orthodontics and Dentofacial Orthopedics : Official Publication of the American Association of Orthodontists, Its Constituent Societies, and the American Board of Orthodontics*, **130**(4), 523–530. DOI : 10.1016/j.ajodo.2005.10.023.

SOLLACI L. B. & PEREIRA M. G. (2004). The introduction, methods, results, and discussion (imrad) structure : a fifty-year survey. *Journal of the Medical Library Association*, **92**(3), 364–371.

SWALES J. (1981). *Aspects of Article Introductions*. Language Studies Unit, University of Aston in Birmingham. Google-Books-ID : Gok7NAAACAAJ.

SWALES J. M. (1990). *Genre analysis*. Cambridge university press.

TEUFEL S., CARLETTA J. & MOENS M. (1999). An annotation scheme for discourse-level argumentation in research articles. In H. S. THOMPSON & A. LASCARIDES, Éds., *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, p. 110–117, Bergen, Norvège : Association for Computational Linguistics.

THOMPSON D. K. (1993). Arguing for experimental "facts" in science : A study of research article results sections in biochemistry. *Written Communication*, **10**(1), 106–128. DOI : 10.1177/0741088393010001004.

TOKALA Y. S. S. S., ALURU S. S., VALLABHAJOSYULA A., SANYAL D. K. & DAS P. P. (2023). Label informed hierarchical transformers for sequential sentence classification in scientific abstracts. *Expert Systems*, **40**(6), e13238. DOI : 10.1111/exsy.13238.

VINKERS C. H., TIJDINK J. K. & OTTE W. M. (2015). Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014 : retrospective analysis. *BMJ*, **351**, h6467. DOI : 10.1136/bmj.h6467.

YAMADA K., HIRAO T., SASANO R., TAKEDA K. & NAGATA M. (2020). Sequential span classification with neural semi-markov crfs for biomedical abstracts. In T. COHN, Y. HE & Y. LIU, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 871–877, En ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.77.

ŠAUPERL A., KLASINC J. & LUZAR S. (2008). Components of abstracts : Logical structure of scholarly abstracts in pharmacology, sociology, and linguistics and literature. *JASIST*, **59**, 1420–1432. DOI : 10.1002/asi.20858.

Annexes

A Constitution de la taxonomie de revendications

PubMed	AZ-CL	ART	DRI	MuLMS-AZ	Description
OBJECTIVE	AIM	HYPOTHESIS MOTIVATION GOAL	CHALLENGE	MOTIVATION	A sentence describing the research target, goal, aim or the motivation for the research.
BACKGROUND	BACKGROUND CONTRAST BASIS	BACKGROUND	BACKGROUND	BACKGROUND PRIORWORK	A statement concerning the knowledge domain or previous related work.
METHOD	OWN	OBJECT, METHOD MODEL EXPERIMENT OBSERVATION	APPROACH	EXPERIMENT PREPARATION CHARACTERIZ. EXPLANATION	A sentence describing the research procedure, models used, or observations made during the research.
RESULT	OWN	RESULT	OUTCOME	RESULTS EXPLANATION	A sentence describing the study findings, effects, consequences, and/or analysis of the results.
CONCLUSION	OWN	CONCLUSION	OUTCOME FUTUREWORK	CONCLUSION	A statement concerning the support or rejection of the hypothesis or suggestions of future research.
–	TEXT OTHER	–	SENTENCE UNSPECIFIED	–	Example sentences, broken sentences, etc.

FIGURE 4 – Inventaire et alignement de plusieurs taxonomies de Zonage Argumentatif (directement repris de (Schrader *et al.*, 2023)). La taxonomie AZ-CL est celle de Teufel *et al.* (1999), élaborée à partir d'articles présentés à des conférences ou des ateliers affiliés à ACL.

phase	#categ.	anno.	#phrases	#articles	$\alpha(\uparrow)$	$\kappa(\uparrow)$ (min-max)
1	5	a1, a2, a3, a4	987 (a1,2) ; 246 (a3,4)	10 (a1-2) ; 4 (a3,4)	0,58	0,09-0,70
2	5	a1, a2, a5, a6	176	2	0,67	0,62-0,73
3	8	a1, a2	622	4	0,57	0,57
4	8	a1, a2	289	2	**0,81**	**0,81**

TABLE 4 – Statistiques des différentes phases d'annotation. En tout, six annotateur·ices (deux chercheur·euses, trois doctorant·es et une étudiante de Master en TAL) ont pris part à la campagne. α : alpha de Krippendorff (Krippendorff, 2011, 2013), κ : kappa de Cohen (Cohen, 1960).

B Performance des modèles et distribution des prédictions

modèle	CONT.	CONTR.	RÉS.	IMP.	DIR.	LIM.	PLAN	NR	F1 moyen
LR+cv+text	0.50	0.40	0.57	0.21	0.26	0.07	0.54	0.66	0.54
LR+cv+prefix_text	0.69	0.51	0.64	0.11	0.45	0.15	**0.72**	0.79	0.66
LR+cv+prefix_SEP	0.64	0.43	0.58	0.09	0.24	0.18	0.34	0.59	0.54
LR+tf+text	0.49	0.41	0.57	**0.24**	0.25	0.09	0.43	0.66	0.54
LR+tf+prefix_text	0.68	0.50	0.64	0.21	0.39	0.12	0.68	0.78	0.66
LR+tf+prefix_SEP	0.67	0.49	0.64	0.21	0.46	0.12	0.67	0.78	0.65
SVM+cv+text	0.54	0.41	0.64	0.00	0.40	0.10	0.54	0.72	0.59
SVM+cv+prefix_text	**0.70**	**0.57**	0.64	0.11	0.47	0.18	0.67	**0.81**	**0.68**
SVM+cv+prefix_SEP	0.69	0.53	0.58	0.11	**0.48**	0.18	0.62	0.79	0.66
SVM+tf+text	0.52	0.41	0.63	0.22	0.40	0.11	0.61	0.72	0.59
SVM+tf+prefix_text	0.66	0.53	**0.67**	0.21	0.47	**0.19**	0.62	**0.81**	**0.68**
SVM+tf+prefix_SEP	0.67	0.54	**0.67**	0.21	0.47	0.18	0.65	**0.81**	**0.68**

TABLE 5 – F1-scores moyens des modèles de Régression Logistique (LR) et de Machines à Vecteurs de Support (SVM), avec des vecteurs de sacs de mots (`cv`) et TF-IDF (`tf`), et plusieurs configurations pour l'entrée : phrase seule (`text`), avec préfixe de section (`prefix`), avec section séparée par un séparateur spécial (`prefix_SEP`).

modèle	CONT.	CONTR.	RÉS.	IMP.	DIR.	LIM.	PLAN	NR	F1 moyen
RoBERTa+prefix_SEC	0.84	0.79	0.80	0.30	0.69	0.46	0.79	0.91	0.83
RoBERTa+prefix_SEC_lr	0.88	**0.87**	0.86	0.37	0.78	0.58	0.87	0.94	0.88
RoBERTa+prefix_SEC_ll	0.90	0.84	0.85	0.56	0.77	0.51	0.85	0.93	0.87
DeBERTa+prefix_SEC	0.81	0.76	0.82	0.00	0.73	0.39	0.85	0.92	0.82
DeBERTa+prefix_SEC_lr	0.90	0.85	0.85	0.48	0.81	**0.54**	0.84	0.94	0.88
DeBERTa+prefix_SEC_ll	0.91	0.86	**0.87**	0.45	**0.83**	0.60	**0.90**	0.94	**0.89**
SciBERT+prefix_SEC	0.86	0.81	0.80	**0.57**	0.74	0.44	0.82	0.92	0.84
SciBERT+prefix_SEC_lr	**0.93**	0.86	**0.87**	0.48	0.80	**0.54**	0.85	0.94	**0.89**
SciBERT+prefix_SEC_ll	**0.93**	**0.87**	0.86	0.52	0.82	0.51	0.88	**0.95**	**0.89**

TABLE 6 – F1-scores moyens des différents modèles basés sur BERT, avec plusieurs configurations de contexte ajouté à la phrase cible dans l'entrée : l'intitulé de section (`prefix_SEC`), la phrase précédente et la phrase suivante (`lr`), les deux phrases précédentes (`ll`).

	#anno.	CONT.	CONTR.	PLAN	RÉS.	IMP.	DIR.	LIM.	NR
annoté	15 992	14,1 %	12,7 %	2,4 %	22,3 %	1,0 %	2,8 %	3,5 %	41,2 %
ensemble	16 126 896	8,9 %	6,9 %	0,9 %	15,7 %	0,7 %	1,6 %	1,7 %	63,6 %

TABLE 7 – Nombre total d'étiquettes et part de chaque catégorie dans les annotations manuelles et dans l'ensemble du corpus. Les multi-étiquettes sont comptées une fois dans chacune des catégories concernées.

C Exemples issus du corpus

catégorie	exemple issu du corpus (traduit de l'anglais)
CONTEXTE	« Quelle que soit l'approche finalement adoptée [...] peu d'attention est portée en général sur les variantes de prononciation issues des processus d'enchaînement de la parole, de l'hypoarticulation et d'autres phénomènes typiques du discours familier [...] » (Lukeš *et al.*, 2018)
CONTRIBUTION	« Notre architecture est une variante du modèle Seq2Seq où deux décodeurs distincts sont utilisés au lieu d'un seul dans l'architecture originale. » (Dinarelli & Grobol, 2019)
PLAN	« Dans la Section 2, nous donnerons un aperçu des avantages principaux de cette approche. » (Och & Ney, 2001)
RÉSULTAT	« Nous voyons dans ce graphique que le classement relatif des modèles reste le même, sauf pour les tailles 1-3 où le parseur probabilistique fait mieux (ou aussi bien) que les modèles basés sur des classificateurs non lexicalisés. » (Eryiğit *et al.*, 2008)
IMPACT	« Nous croyons qu'il s'agit d'un moment critique dans la vie de la recherche sur les systèmes de dialogue, et nous anticipons d'excitantes avancées dans un futur proche, menant à des systèmes non seulement utiles mais également simples d'utilisation [...] » (Glass & Seneff, 2003)
DIRECTIONS	« Une bonne métrique devrait utiliser toute l'information que nous avons, et notamment les arbres de référence, pour évaluer. » (Headden III *et al.*, 2008)
LIMITATION	« [...] le processus de génération de ré-écritures à l'aide de LLMs peut être coûteux en calculs, nécessitant des ressources importantes en GPU et pouvant prendre plusieurs heures pour de grands jeux de données. » (Fan *et al.*, 2023)
NON REVENDICATION	« Nous avons aléatoirement sélectionné un jeu de données de 150 tweets annotés par les deux annotateurs en catégories grammaticales et en structures de dépendances. » (Bhat *et al.*, 2018)

TABLE 8 – Sélection aléatoire de revendications (également une phrase NON REVENDICATION) du corpus (prédictions effectuées par le modèle SciBERT).

catégorie	exemple issu du corpus (traduit de l'anglais)
RÉSULTAT +LIMITATION	« Ceci indique que l'insertion de nouveaux tours impliquant une connaissance supplémentaire du domaine pourraient interrompre les dialogues originaux, ce qui complique le processus de dialogue et rend la modélisation des tours originaux difficile. » (Gao *et al.*, 2022)
RÉSULTAT +CONTRIBUTION	« Dans cet article, nous proposons une nouvelle méthode d'appariement et d'extraction de *TM* basée sur le *Universal Sentence Encoder* (Cer et al., 2018), capable d'identifier des segments sémantiquement similaires dans les *TMs* mieux que les méthodes basées sur la distance d'édition. » (Ranasinghe *et al.*, 2020)
IMPACT +DIRECTIONS	« Ceci pourrait encourager la création de modèles pour des langues moins dotées et moins parlées, même si de tels modèles n'ont dans l'immédiat pas un taux d'erreur de mots assez bas pour faire concurrence aux modèles pour l'anglais. » (Merz & Scrivner, 2022)

TABLE 9 – Sélection aléatoire de revendications multi-étiquettes du corpus (prédictions effectuées par le modèle SciBERT).

score de certitude	**exemple issu du corpus (traduit de l'anglais)**
2.14	« Cependant, il reste encore beaucoup à faire dans des recherches futures, car le modèle le plus performant sur la tâche partagée *Fever* atteint un score de précision de 68,21 % (Thorne et al., 2018b) » (Lee *et al.*, 2020)
3.48	« En réponse à la question, "Dans quelle salle a lieu la réunion de lundi ?", il peut n'y avoir aucun moyen de choisir entre "A101" et "A201" sans connaissance complémentaire du contexte. » (Jamison & Gurevych, 2013)
5.97	« Pour le traçage vers le bas, il n'y a aucun problème avec les règles de suppression utilisant "0" ou "[]", qu'elles soient conditionnées par un environnement ou non, et le traçage de l'application vers le bas des règles d'épenthèse avec un environnement ne pose pas non plus de problème. » (Kilbury *et al.*, 2011)

TABLE 10 – Phrases du corpus présentant un score de certitude minimal, intermédiaire, et maximal.

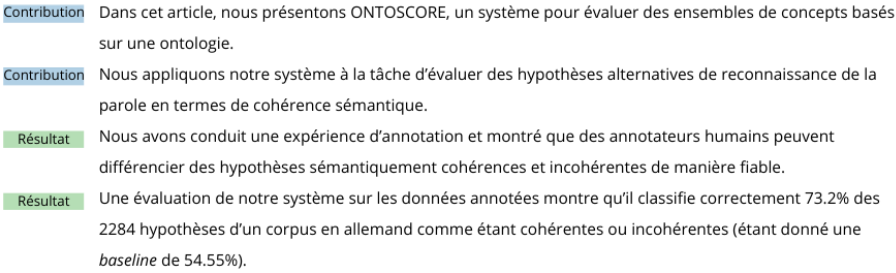

FIGURE 5 – Exemple d'un résumé dont la structure en CONTRIBUTION-RÉSULTAT est typique des articles du corpus publiés entre 1995 et 2009. Résumé de Gurevych *et al.* (2003), traduit en français.

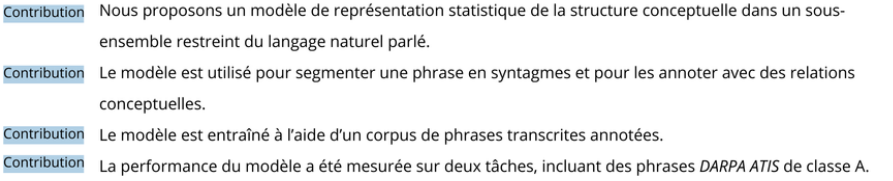

FIGURE 6 – Exemple d'un résumé dont la structure en CONTRIBUTION est typique des articles du corpus publiés avant 1995. Résumé de Pieraccini *et al.* (1991), traduit en français.

D Structure des résumés issus de l'Anthologie ACL *vs.* ArXiv

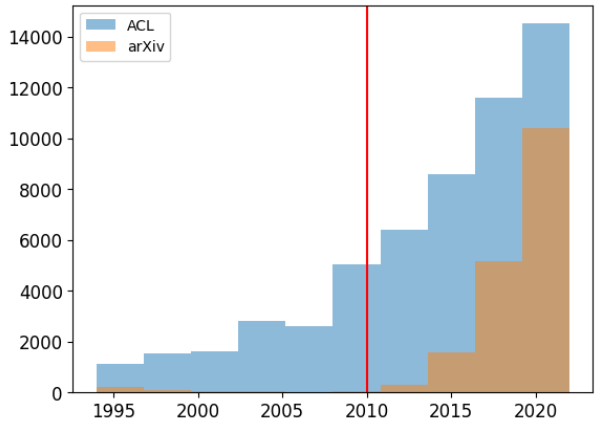

FIGURE 7 – Histogrammes du nombre d'articles issus de l'Anthologie ACL et d'ArXiv par année, sur la période commune 1952-2022. Les distributions sont davantage similaires entre 2010 et 2022 qu'avant, ce qui permet la comparaison des structures de résumés sur cette période par des diagrammes de Sankey.

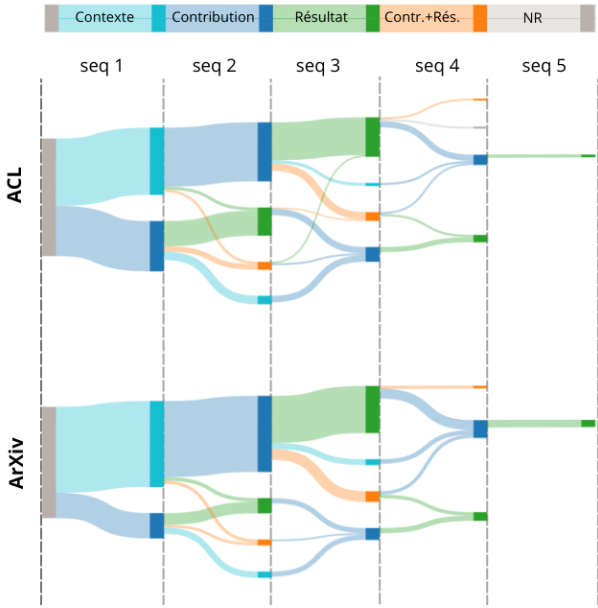

FIGURE 8 – Diagrammes de Sankey permettant de visualiser la distribution des structures de résumés issus de l'Anthologie ACL *vs.* ArXiv, sur la période 2010-2022. L'interruption d'un flux indique la fin de la structure. Seuls les flux identifiés dans plus de 1 % des résumés sont représentés.