Analyse des *claims* dans les articles de Traitement Automatique des Langues à l'aide d'une méthode par apprentissage non supervisé

Mémoire de Master 1 Langue et Informatique

Présenté par :
Fanny Ducel

Sous la direction de :
Karën FORT (Sorbonne Université, LORIA)

Table des matières

Table des figures

Liste des tableaux

Remerciements

Je souhaiterais remercier chaleureusement différentes personnes qui m'ont aidée tout au long de ce travail de mémoire.

Tout d'abord, un grand merci à mes encadrant·es Karën Fort et Maxime Amblard qui m'ont proposé ce projet, m'ont conseillée et m'ont accordé beaucoup de temps ces six derniers mois.

Je tiens également à remercier Gaël Lejeune et Carlos González pour leurs idées et leurs explications sur certains points techniques concernant le *clustering*, mais aussi Patrick Paroubek qui a pris le temps de discuter avec moi de son outil, et Cédric Wemmert pour ses éclairants *e-mails* sur l'évaluation du *clustering*. Merci également à Aurélie Névéol pour ses retours sur le manuscrit.

Merci beaucoup à ma promotion de M1, et particulièrement à Peng pour les corrections qu'il a apportées à mon corpus de prénoms chinois.

Enfin, un grand merci à la direction du LORIA, et plus précisément au financement de CODEINE, grâce auquel j'ai pu effectuer un stage pour finir ce mémoire dans les meilleures conditions possibles, et à toutes les personnes du bureau 230 pour leur accueil chaleureux et, pour certain·es (Hee-Soo, Priyansh, Vincent), leur aide précieuse et leur partage de connaissances.

Introduction

Motivations

Les ordinateurs peuvent-ils comprendre le langage naturel ? Le test de Turing a-t-il été réussi ? Si l'on en croit certains titres d'articles de presse [1], on pourrait croire que oui. Nous savons que la réalité est tout autre [Bender and Koller, 2020]. Alors, comment se fait-il que de telles affirmations apparaissent dans des articles de presse scientifique ?

Il s'avère que les journalistes s'appuient sur les *claims*, c'est-à-dire les affirmations tirées des résultats et mises en avant dans les articles de recherche.

Par exemple, on peut trouver des phrases comme : « Les travaux récents [...] ont donné des résultats prometteurs en matière de modélisation du langage. » [2] ou « Nous démontrons également des situations spécifiques et courantes dans lesquelles la SHDGA se heurtera invariablement à une inefficacité et un non-déterminisme graves, et que l'EAA traitera de manière efficace et déterministe. » [3]

Or, produire des *claims* avec un degré de certitude trop élevé par rapport à la réelle significativité des résultats pose problème, notamment d'un point de vue éthique. Les résultats sont diffusés avec une amplification injustifiée, ce qui peut pousser la communauté à croire que certaines tâches ont été achevées avec grand succès alors que ce n'est pas le cas. On peut également opter pour l'utilisation d'un système dont les performances ont été surestimées, ce qui peut entraîner des baisses de résultats sur nos propres travaux.

Cette surproduction de *claims* et la survalorisation de son travail peut s'expliquer par le système actuel du « publier ou périr » (de l'expression anglaise *« publish or perish »*, utilisée pour la première fois par Case [1927]), qui pousse les membres de la communauté de recherche scientifique à publier un maximum d'articles, et donc à parfois préférer la quantité à la qualité.

Ces phénomènes de surévaluation des résultats sont indésirables, voire condamnables, d'autant plus qu'ils s'inscrivent dans un contexte de méfiance envers les scientifiques qu'il faudrait au contraire chercher à limiter.

Cette étude aborde ainsi des enjeux éthiques. L'éthique devient en effet un aspect crucial du TAL, car la popularité du domaine entraîne une utilisation massive et quotidienne des applications utilisant du TAL ; les conséquences ne sont alors pas négligeables. C'est notamment ce que mettent en lumière Jin et al. [2021] ainsi que Hovy and Spruit [2016] en abordant l'impact social du domaine et le « bien social » qu'il peut engendrer ou non.

1. Voir par exemple *« Machines That Can Understand Human Speech : The Conversational Pattern Of AI »* https://www.forbes.com/sites/cognitiveworld/2020/06/28/machines-that-can-understand-human-speech-the-conversational-pattern-of-ai/ et *« Computer AI passes Turing test in 'world first' »* *https://www.bbc.com/news/technology-27762088*

2. *« Recent work [...] have shown promising results on language modeling. »*[Zhang and Song, 2019]

3. *« We also demonstrate specific and common situation in which SHDGA will invariably run into serious inefficiency and nondeterminism, and which EAA will handle in an efficient and deterministic manner.»* [Martinovic and Strzalkowski, 1992].

Ce travail s'inscrit également dans le domaine du « TAL pour le TAL » (*NLP4NLP*, d'après [Mariani et al., 2019]), qui consiste en l'utilisation d'outils de TAL sur des articles de TAL.

Enfin, l'étude et l'analyse des *claims* dans les articles scientifiques entrent également dans les champs de l'identification de la subjectivité et de la vérification de faits (*fact checking*).

Définitions

Pour aborder plus en détails ce sujet, nous avons recours à plusieurs notions empruntées à la terminologie linguistique, ainsi qu'à un terme issu du domaine de l'informatique. Elles sont détaillées plus amplement dans l'état de l'art (voir Chapitre 1) mais nous les définissons une première fois ici.

Tout d'abord, les notions linguistiques de *claim* et de modalité épistémique sont centrales.

D'après Blake [2010], un *claim* est « une affirmation qui rend compte de quelque chose qui entraîne un effet ou un résultat » [4].

Ces *claims* sont exprimés avec différents degrés de certitude, et s'avèrent notamment problématiques quand ils paraissent trop certains, ou au contraire, trop incertains. En linguistique, plutôt que de parler d'expression de la certitude, on fait appel à la notion de *modalité épistémique*. Il s'agit de « l'expression linguistique d'une estimation de la probabilité qu'un état de choses particulier soit, ait été ou devienne vrai »[Rubin, 2006] [5].

La modalité épistémique inclut et est concrètement réalisée par les *hedges*, qui sont alors très présents dans nos *claims*. Ce sont des expressions qui rendent les messages indéterminés, inexacts ou atténués et les présentent comme des opinions plutôt que des faits (définition créée à partir de [Martín-Martín, 2008] et [Hyland, 1998]). C'est par exemple le cas de verbes, qu'ils soient modaux ou non : « pouvoir », « devoir », « sembler », « prouver », ... Certains adjectifs et adverbes peuvent également remplir ces fonctions : « probable », « évident », « potentiellement », « indéniablement », ...

L'utilisation d'*hedges* pour renforcer ou diminuer le degré de certitude d'un *claim* peut contribuer à produire des affirmations inappropriées, dont l'intensité ne coïncide pas avec les résultats réellement obtenus. Ce phénomène de production d'affirmations inappropriées est nommé « embellissement » (traduction du terme anglais *« spin »*, proposée par Koroleva [2017]) dans le domaine du traitement automatique des langues et est donc issu d'une terminologie informatique.

4. *« Claims that capture something that brings about an effect or a result »*
5. *« Certainty, or epistemic modality, is a linguistic expression of an estimation of the likelihood that a particular state of affairs is, has been, or will be true. »*

Réalisations

La détection de l'embellissement est une tâche difficile, à la fois sur le plan technique et sur le plan théorique car elle nécessite de savoir à partir de quel score on peut estimer qu'un résultat est réellement satisfaisant et justifie l'emploi de tel ou tel mot. Notre travail se concentre donc uniquement sur la détection des *claims* et leur classification selon leur modalité épistémique, ce qui constitue une première étape essentielle pour une éventuelle future étude sur l'embellissement en lui-même.

Nous avons décidé d'effectuer notre étude sur un corpus en anglais contenant 6 372 articles publiés entre 1979 et 2020, issus de la conférence CORE A* du domaine du Traitement Automatique des Langues (TAL) : Association for Computational Linguistics (ACL).

Nous avons opté pour un système de classification non-supervisée, avec du *clustering*, car l'annotation manuelle se serait révélée peu fructueuse. Notre état de l'art (voir Section 1.1) prouve en effet que l'annotation de *claims* et d'*hedges* est très chronophage et résulte en des accords inter-annotateurs relativement bas.

De ce fait, notre étude est celle qui utilise le corpus de plus grande envergure, puisque la nécessité d'annotation manuelle cantonnait les autres travaux à utiliser des corpus de quelques dizaines d'articles (à l'exception du corpus d'un peu plus de 1 000 documents de Li et al. [2017] et de celui de presque 4 000 articles de Koroleva [2017]).

A notre connaissance, il s'agit également de la première étude portant sur les *claims* présents dans les articles de TAL puisque les autres travaux sur ce sujet ont été réalisés sur des corpus d'articles biomédicaux ou journalistiques.

Ainsi, nous avons extrait les *claims* présents dans les différentes sections des articles, puis avons utilisé des techniques de *clustering* pour les classifier en trois catégories selon leur degré de certitude.

A partir de ces résultats, nous avons ensuite pu analyser et comparer l'utilisation des différentes catégories de *claims* selon différents facteurs.

Finalement, nous avons étendu nos comparaisons et analyses à d'autres approches afin de voir si le genre des auteur·ices, leur continent d'origine et leur affiliation à des institutions prestigieuses jouent un rôle dans leur utilisation des *claims*.

Chapitre 1

État de l'art

Cet état de l'art nous permet d'aborder plus en détails les notions linguistiques précédemment évoquées de *claims*, d'*hedging* et de modalité épistémique. Nous réalisons un diagramme de Venn afin d'illustrer l'imbrication entre ces différentes idées : les *claims* peuvent contenir des marques de modalité épistémique, et certaines de ces marques sont plus précisément des *hedges*.

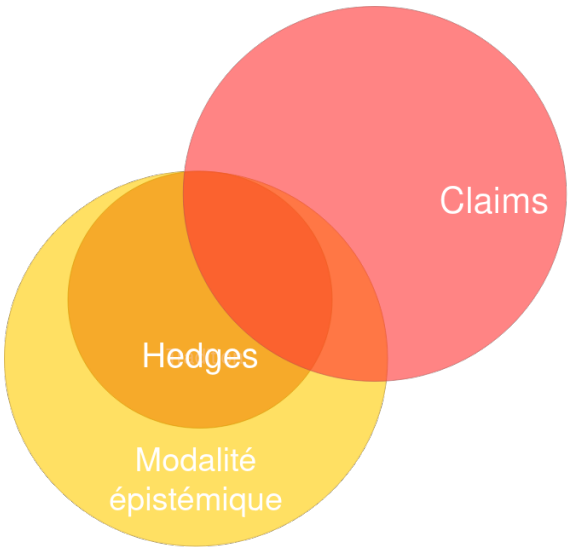

FIGURE 1.1 – Diagramme de Venn illustrant les notions de *claims*, *hedges* et modalité épistémique

1.1 Détection de claims

Comme mentionné dans l'introduction, la tâche de détection de *claims* dans des articles scientifiques a déjà été réalisée, mais, à notre connaissance, uniquement sur des articles scientifiques de biomédecine et des articles de presse, et en utilisant de la classification purement manuelle, ou automatique supervisée, qui est alors dans la plupart des cas une classification binaire (une phrase est ou n'est pas un *claim*).

La première étude que nous avons trouvée à ce sujet est celle de Martín-Martín [2008]. L'auteur donne des raisons d'utilisation de l'*hedging* et nous permet ainsi de mieux comprendre ce phénomène. Selon lui, les claims contenant de l'*hedging* sont utilisés pour réduire le risque d'opposition et maximiser les chances qu'un article soit reconnu au sein de la communauté scientifique. Il pointe également du doigt la dimension rhétorique et sociale du discours académique.

Cependant, il nuance ses propos en citant également Salager-Meyer, pour qui l'utilisation de *claims* vagues est positive et possède de réels intérêts. Cette imprécision pourrait en effet permettre de « montrer au lectorat que l'on n'a pas le dernier mot sur le sujet » [1] et s'inscrire dans des valeurs essentielles en science : l'incertitude, le scepticisme et le doute. Contrairement aux intuitions que nous avions, les *claims* imprécis pourraient ainsi augmenter la crédibilité d'une affirmation et d'un·e scientifique.

Martín-Martín [2008] crée une classification des différentes stratégies utilisées dans les allégations scientifiques :

— la stratégie d'indétermination : traduite par la modalité épistémique (verbes modaux, (semi-)auxiliaires, verbes exprimant la possibilité, adverbes modaux, noms et adjectifs modaux) et les approximateurs de quantité/fréquence/degré/temps
— la stratégie de subjectivisation : traduite par l'utilisation de pronoms de première personne suivis de verbes de cognition ou verbes performatifs, l'utilisation de formules comme « to our knowledge », « in our view », « in my experience » (*à notre connaissance, de notre point de vue, selon mon expérience*), ...
— la stratégie de dépersonnalisation : constructions impersonnelles actives et passives (sans agent exprimé)

Il analyse ensuite empiriquement un corpus de 40 articles (20 en anglais et 20 en espagnol) puis calcule les pourcentages d'utilisation de chacune de ces stratégies dans chaque partie des articles. Il en conclut que les anglophones utilisent plus la stratégie d'indétermination, « atténuant ainsi la force de leurs affirmations dans le but d'obtenir une plus grande acceptation de la part des membres de la communauté de recherche » [2], mais que dans les deux cas on retrouve le plus de *claims* dans les introductions et les parties discussions/conclusions.

L'étude de Martín-Martín [2008] nous permet ainsi de nous intéresser aux raisons qui poussent les auteur·ices à écrire des *claims* imprécis. De plus, ses expériences nous donnent une première idée sur les parties des articles qui contiennent le plus de *claims* ainsi que sur les pays d'origine des auteur·ices qui en écrivent le plus.

Blake [2010] définit quant à elle cinq manières de communiquer ses conclusions : affirmations explicites, affirmations implicites, corrélations, comparaisons et observations.

Elle décrit en détails son processus d'annotation par phrase, pour lequel elle a ensuite développé une interface avec ses étiquettes. Cela lui permet de constater que les *claims* sont les plus présents dans la section Discussion, puis Introduction, puis Résultats. En mettant en place ce guide d'annotation précis, elle parvient à obtenir un kappa de Cohen de 0,88 (notons toutefois que ce score ne concerne pas la classification interne des *claims explicites*, notre sujet d'étude, mais seulement de phrases étant ou non des *claims*).

Elle développe plusieurs approches automatiques. L'une est basée sur des grammaires de dépendances, l'autre sur des caractéristiques plus linguistiques qui lui permettent notamment d'extraire les affirmations explicites à l'aide d'arbres de dépendance lexico-syntaxique. Les résultats de son étude, sur 29 articles complets, montrent que les auteur·ices « rapportent généralement des affirmations explicites (77,12 %) plutôt que des observations (9,23 %), des corrélations (5,39 %), des comparaisons (5,11 %) ou des affir-

1. *« She also argues that academics may choose to remain vague in their claims to show their readers that they do not have the final word on the subject. »*

2. *« The English-speaking writers resort more frequently to making their claims more tentative and indeterminate, and thus mitigate the strength of their assertions in a bid to achieve greater acceptance from the members of the research community. »*

mations implicites (2,7 %) »[3]. Ils lui permettent également de prouver l'importance de prendre en compte la totalité du contenu des articles, et pas seulement les résumés, car seuls 7,84 % des *claims* y sont explicités.

Cette étude est pertinente pour notre travail pour plusieurs raisons. Tout d'abord, Blake [2010] prouve que l'annotation requiert un travail de préparation conséquent. Par la suite, ses résultats indiquent que ce sont les affirmations explicites qui sont les plus utilisées. Sa définition d'affirmations explicites se rapproche énormément de notre définition de *claims* et suggère que notre étude sera en mesure de couvrir la majorité des conclusions que tirent les auteur·ices. Finalement, ses remarques sur l'importance de ne pas s'en tenir seulement aux résumés des articles nous est d'une grande aide puisque notre hypothèse de départ était qu'il s'agirait au contraire de la partie contenant le plus de *claims*. Nous aurions pu nous intéresser seulement à cette partie, ce qui aurait grandement limité nos analyses.

Koroleva [2017] écrit le seul article en français de cet état de l'art, dans lequel elle fait un travail de détection d'affirmations inappropriées dans 3 938 publications scientifiques biomédicales. Elle se concentre sur les « affirmations où l'effet positif du traitement est plus grand que celui effectivement prouvé par la recherche ». Elle constate que ces affirmations sont surtout présentes dans les résumés et en explique les conséquences directes : les médecins auront tendance à accorder plus d'efficacité aux articles contenant du *spin*, ce qui peut affecter leur prise de décision lors du choix d'un traitement. De même, cela pose problème quand ces résumés sont repris dans des communiqués de presse et dans des articles d'actualité de santé. Cela est d'autant plus vrai que, parfois, seuls les résumés sont gratuitement accessibles en ligne, ce qui « augmente ainsi fortement l'impact du *spin* sur la diffusion des résultats de recherche ».

Elle effectue sa propre classification des affirmations inappropriées :

— présentation inappropriée des résultats de recherche : les effets négatifs et/ou certains résultats sont omis, le contexte de l'étude et/ou du corpus est imprécis, on utilise du *spin* linguistique (notamment avec des mots évaluatifs positifs forts)
— interprétation inappropriée des résultats de recherche : une affirmation est faite alors que les résultats n'ont pas de significativité statistique ni de pertinence clinique, qu'aucun test comparatif n'a été mené et l'essai n'est pas randomisé
— extrapolation inappropriée : on présente une population plus large, une intervention ou un résultat différent de ceux effectivement évalués

Son étude est néanmoins plus large que la nôtre et moins focalisée sur les *claims*, car elle extrait l'évaluation du traitement, les entités, les paraphrases, et fait un point sur les biais présents. Elle admet que l'absence de corpus annoté empêche la mise en place d'un apprentissage automatique et décide donc de créer d'abord un système à base de règles. Pour cela, elle identifie manuellement les phrases du corpus qui contiennent des mots pouvant être utilisés pour décrire les résultats obtenus et en crée des règles à bases de grammaires locales (*via* `Unitex`). Elle poursuit plus ou moins ce travail dans sa thèse [Koroleva, 2020] ainsi que dans un autre article dans lequel elle présente un outil de détection de *spin* dans les publications biomédicales [Koroleva et al., 2020].

Puis, Li et al. [2017] mènent une étude sur les affirmations exagérées dans les actualités scientifiques, en utilisant 462 communiqués de presse liés à la santé et leurs *claims*

3. *« The results also show that authors typically report explicit claims (77.12 %) rather than an observations (9.23 %), correlations (5.39 %), comparisons (5.11 %) or implicit claims (2.7 %) »*

correspondants dans 668 articles journalistiques. Toutes ces données sont manuellement annotées, la classification sera donc supervisée. Ils commencent par envisager sept classes de relation à l'intérieur des *claims* puis n'en retiennent que quatre : pas de relation, corrélation, causalité conditionnelle, et causalité. Plusieurs représentations sont testées : simple sac de mots, sacs de mots enrichis grâce à des indices linguistiques définis manuellement ou grâce à des analyses syntaxiques de dépendance. Plusieurs algorithmes de classification supervisée sont également testés, mais les meilleurs F1-score (qui s'élèvent à 0,718 dans les articles de journaux et 0,607 dans les communiqués de presse) sont obtenus avec le SVM.

Leurs expériences nous présentent ainsi les avantages et les inconvénients liés à l'apprentissage supervisé. Néanmoins, l'inconvénient principal réside dans la taille du corpus, qui se voit limitée par la nécessite d'annotation manuelle, et entre en conflit avec le corpus que nous souhaitons analyser. L'un des enjeux de leur étude coïncide avec l'un des nôtres. En effet, ils cherchent à prouver que les médias ont tendance à amplifier les *claims*, d'autant plus quand ils sont surévalués dès la publication scientifique.

Le travail de Luttenberger and Vulinovic [2018] est assez semblable puisque leur objectif est d'« identifier la force des affirmations pour détecter les exagérations dans les actualités scientifiques » [4], et ce en comparant un corpus journalistique avec un corpus d'articles scientifiques. La catégorisation utilisée est presque identique à celle de Li et al. [2017], et les meilleurs résultats sont également obtenus avec un SVM, en utilisant des sacs de mots et en conservant les mots outils.

Nous retenons des conclusions semblables à celles que nous avons tirées de l'article de Li et al. [2017]. Toutefois, ce travail nous permet de remettre en question la pertinence du filtrage des mots outils. Nous gardons cela en tête pour la suite de notre travail et décidons donc de mener des expériences de *clustering* en gardant ces mots.

Finalement, Patro and Baruah [2021] proposent une approche pour détecter automatiquement les exagérations dans les affirmations présentes dans les actualités de santé. Ils comparent alors par paires de déclarations, l'une issue d'un article scientifique et l'autre de l'article de presse dérivé. Ils utilisent sept catégories (semblables à celles initialement envisagées par Li et al. [2017]) et entraînent des classifieurs afin de pouvoir estimer dans quelle mesure les écrits journalistiques exagèrent les affirmations des articles scientifiques.

Tout comme les deux études précédentes, celle-ci permet à nouveau de prouver l'écart entre les *claims* explicités dans les articles scientifiques et ceux que l'on retrouve dans les journaux. Cela confirme que les enjeux de notre étude sont importants.

Test de l'outil `DeSpin`

Nous avons souhaité utiliser les outils `DeSpin` et `SynDepng` développés par P. Paroubek et A. Koroleva et présentés notamment dans [Koroleva et al., 2020]. Cependant, ces systèmes ont été pensés pour les articles du domaine biomédical, pour une approche syntaxique et dans un but d'annotation.

`DeSpin` compte une série de fonctionnalités liés à la détection de la structure, des *hedges* et des résultats obtenus. Cependant, ces fonctionnalités dépendent de librairies

4. Le titre de leur publication est : *« Claim Strength Identification for Detecting Exaggerations in Science News »*

Python qui ont entre temps été mises à jour, les scripts ne sont donc plus fonctionnels. Nous avons pu échanger avec P. Paroubek qui nous a indiqués que l'actualisation de l'outil est actuellement en cours mais ne sera pas disponible à temps pour que nous puissions véritablement utiliser le logiciel dans le cadre de ce mémoire.

Toutefois, nous avons pu avoir accès aux codes et nous avons remarqué qu'ils reposent également sur beaucoup de méthodes de *substring matching* qui utilisent des listes d'indices propres au domaine du biomédical avec des tokens comme *patient, symptôme, traitement, médicament, neuroleptique, ...*

En parallèle, `DeSpin` permet d'extraire les résultats obtenus dans les articles et de comparer ceux qui sont annoncés dans le résumé et ceux qui sont dans le reste de l'article en utilisant des mesures de similarité. C'est ainsi le phénomène de *spin* qui est détecté. Néanmoins, la méthode d'extraction des résultats obtenus repose sur l'utilisation du numéro d'enregistrement de l'essai clinique (*trial registry number*) et la récupération externe des résultats décrits dans cet enregistrement. Or, nous n'avons pas accès à une telle ressource pour le TAL, il faudrait donc trouver une autre manière d'extraire les résultats énoncés.

Leur système de détection d'*hedges* se base quant à lui sur la liste de tokens suivante : *may, might, can, could, would, seem, appear, suggest, potentially, possibly, potential, possible*. Nous retrouvons ainsi, à l'exception de *would*, des verbes présents dans la liste que nous avons utilisée.

Cela confirme que notre méthode est pertinente puisqu'elle ressemble beaucoup à celle utilisée dans d'autres outils. Celle de P. Paroubek et A. Koroleva est toutefois plus spécifique au domaine qui les intéresse, le biomédical, et nous pouvons donc supposer que leur outil n'aurait pas pu extraire convenablement les résultats décrits dans des articles de TAL ni leur structure. De plus, leur liste d'indices liés à l'*hedging* étant moins conséquente, nous pouvons imaginer que moins d'hedges auraient été détectés.

1.2 Modalité épistémique

Dans le discours scientifiques, les *claims* sont souvent caractérisés par une expression plus ou moins forte de la certitude. En linguistique, cette expression de la certitude est incluse dans la notion de modalité épistémique.

Dans sa thèse, Rubin [2006] se donne pour objectif d'identifier la certitude dans les textes. La modalité épistémique est alors centrale dans son travail, elle considère même que c'est un synonyme de certitude, et la définit comme « l'expression linguistique d'une estimation de la probabilité qu'un état de choses particulier soit, ait été ou devienne vrai » [5]. Elle peut être exprimée par différents moyens lexicaux, sémantiques, syntaxiques et discursifs, notamment les adverbes, les verbes modaux et les phrases à sujet vides, les verbes de réplique, les *tag questions*, les *if-clauses* de condition et les concessions, les marqueurs contrastifs et la passivation.

[Rubin, 2006] lie cette notion aux tâches d'identification de la subjectivité en TAL, et

5. *« Certainty, or epistemic modality, is a linguistic expression of an estimation of the likelihood that a particular state of affairs is, has been, or will be true. »*

affirme que « désormais, la question n'est plus seulement "Qui a fait quoi à qui ?" mais aussi "Qui pense quoi à propos de quelqu'un qui fait quoi ? » [6].

L'autrice réalise par la suite une étude empirique sur 80 reportages et éditoraux journalistiques qui nous donne une première idée de classification de la modalité épistémique. Elle crée en effet son propre modèle de catégorisation de la certitude à cinq étiquettes : certitude absolue, haute certitude, certitude modérée, certitude faible, incertitude.

Elle effectue également un travail d'annotation manuelle, très révélateur de la difficulté de la tâche puisqu'en moyenne, les personnes réalisant l'annotation y ont consacré 10 heures pour 10 articles, et que les accords inter-annotateurs ont atteint au maximum 71 % et le kappa de Cohen s'est élevé à 0,65 au plus (le plus bas étant à 0,13, et la plupart des autres sous 0,5).

D'après ses résultats, elle peut affirmer que l'on trouve 0,82 marqueurs de certitude explicite par phrase dans son corpus et que la catégorie la plus présente est celle de la « haute certitude ». Elle remarque également qu'environ « trois cinquièmes des phrases exprimant de la certitude contiennent un seul marqueur par phrase, un tiers en contient deux, et seulement moins d'un cinquième en contient trois ou plus ». L'utilisation de plusieurs marqueurs dans la même phrase peut avoir tous les usages : renforcer le même degré de certitude ou, à l'inverse, le diminuer ou l'augmenter. Après avoir calculé le niveau de certitude par phrase, elle calcule celui par article. Elle en conclut que les classes grammaticales les plus utilisées comme noyaux de marqueurs de certitude sont les auxiliaires modaux, les adjectifs gradables au superlatif et les intensifieurs adverbiaux.

Ce travail nous fournit donc une définition précise de la notion ainsi que des données chiffrées qui permettent d'une part de témoigner de la complexité de l'annotation manuelle pour cette tâche, et d'autre part d'estimer le nombre de marqueurs de modalité épistémique contenus dans des écrits journalistiques. Nous pourrons ainsi garder à l'esprit ces nombres afin de les confronter à ceux que nous obtiendrons sur des écrits scientifiques de TAL.

Vold [2006] mène une étude semblable, mais sur un corpus d'écrits scientifiques. Elle s'intéresse aux marqueurs de modalité épistémique dans des articles de recherche en linguistique d'une part et en médecine de l'autre. Son approche et ses comparaisons sont également multilingues car les 120 articles de son corpus sont rédigés soit en anglais, soit en français, soit en norvégien. Elle définit la notion comme recouvrant « les expressions linguistiques qui qualifient la valeur de vérité d'un contenu propositionnel » [7] et permettant de décider à quel point on peut se fier à une information, qui peut être donnée comme étant absolument certaine ou absolument incertaine. Elle décide de limiter ces expressions à des unités lexicales et grammaticales, excluant les phrases ou paragraphes entiers.

Elle lie la notion de modalité épistémique à celle d'*hedging*, et émet plusieurs hypothèses quant aux raisons d'utilisation de cette modalité : par humilité, par politesse, pour donner plus de précision au discours, pour anticiper les critiques, ou pour critiquer subtilement les travaux précédents. Ses analyses lui permettent de remarquer que le genre n'a pas d'impact, mais que la langue et la nationalité jouent à l'inverse un rôle important. En effet, les francophones utilisent beaucoup moins de marqueurs de modalité épistémique que les anglophones. Ces derniers font preuve d'une forme « d'hypermodestie » tandis que les francophones de ce que l'on pourrait appeler une « confiance en soi exagérée ».

6. *« The question is no longer just "Who did what to whom ?" but also "Who thinks what about somebody doing what ? »*

7. *« linguistic expressions that qualify the truth value of a propositional content »*

Saurí and Pustejovsky [2012] se basent sur la notion de modalité épistémique pour s'intéresser aux différents degrés de factualité présents dans les textes ainsi qu'au langage spéculatif, et aux notions de certitude et de polarité. Les marqueurs prennent alors différentes formes grammaticales (auxiliaires modaux, adverbes épistémiques, adjectifs, noms, verbes lexicaux), et sont répartis sur une échelle de degrés de factualité qui comptent les étiquettes suivantes : « factuel », « contrefactuel », « probable », « improbable , « possible », « incertain ». Quatre composantes sont prises en compte : l'évènement, la valeur de factualité (liée à la polarité et la modalité épistémique), la source qui assigne la valeur de factualité (l'auteur·ice, *via* des marqueurs), et le temps où la valeur de factualité se passe.

Les différents articles portant sur la modalité épistémique nous permettent alors de mieux aborder cette notion linguistique et de la situer par rapport à l'*hedging*. Quelques catégorisations sont présentées, ce qui nous permet de réfléchir aux étiquettes que nous souhaitons donner à notre système. De plus, des exemples concrets d'expressions de cette modalité sont avancés ; ceux-ci sont en réalité des *hedges*. Nous développons donc la notion d'*hedging* dans la Sous-section suivante.

1.3 *Hedging*

Le phénomène linguistique d'*hedging* est en effet très présent dans les *claims* ; c'est cela qui les rend flous. Il s'agit en effet d'une partie de la modalité épistémique, on peut même considérer que les *hedges* sont l'un des moyens concrets d'exprimer cette modalité. Cette imprécision crée donc parfois une surévaluation des résultats (voire, plus rarement, une sous-évaluation).

Ce terme, dans cette acception, apparaît pour la première fois chez Lakoff [1973b], qui introduit l'idée de degrés de vérité et de « logique floue ». Il souhaite étudier les mots « dont le sens implique implicitement le flou - des mots dont le rôle est de rendre les choses plus ou moins floues » [8]. Il les appelle des « hedges » et en dresse une première liste, contenant essentiellement des adjectifs et des adverbes.

Cette liste sera réutilisée et adaptée à de nombreuses reprises dans des travaux sur la détection de l'*hedging*. C'est le cas d'articles que nous présentons ci-dessous, mais également de notre étude. La liste d'indices que nous avons élaborée pour nos expériences contient en effet des *hedges* initialement listés par Lakoff [1973b].

1.3.1 Définitions proposées

Les différentes études qui ont été menées sur l'*hedging* consacrent souvent une partie à la définition du terme. Nous présentons ici celles qui nous semblent les plus pertinentes.

Tout d'abord, nous définissons ce concept en utilisant les mots de Martín-Martín [2008] : « On entend généralement par *hedging* les expressions du langage qui rendent les messages indéterminés, c'est-à-dire qui transmettent l'inexactitude, ou qui, d'une manière ou d'une autre, atténuent ou réduisent la force des assertions ».

8. *« For me, some of the most interesting questions are raised by the study of words whose meaning implicitly involves fuzziness- words whose job is to make things fuzzier or less fuzzy »*

Cette définition nous permet de comprendre en quoi une utilisation, surtout abusive, de ce procédé peut être problématique dans des articles scientifiques, qui se doivent de partager des données et des résultats fiables et précis.

Hyland [1998] consacre sa thèse à l'*hedging* dans les articles scientifiques de recherche en biologie. En outre de fournir des définitions détaillées, il développe les enjeux recouverts par la notion (voir Sous-section suivante).

L'auteur définit le phénomène d'*hedging* comme une facette de la modalité épistémique qui reflète « un manque d'engagement complet envers la valeur de vérité d'une proposition connexe, ou un désir de ne pas exprimer cet engagement de manière catégorique ».

Il donne également sa définition d'*hedges* : « les moyens par lesquels les auteurs peuvent présenter une proposition comme une opinion plutôt qu'un fait ». Il fournit des exemples concrets de moyens linguistiques reflétant l'objectivité mais aussi une mise à distance de la personne qui écrit avec son contenu, et étudie également la distribution de la modalité dans des articles de biologie : elle est surtout haute en introduction et dans les sections Discussion. Il constate que les *hedges* lexicaux (qu'il estime à plus de 350) constituent la manière la plus commune d'exprimer la modalité épistémique en anglais. Il remarque par ailleurs que ces marques linguistiques disparaissent, et que le degré de certitude des faits augmente quand les études sont reprises dans des magazines populaires. En bref, les *hedges* sont alors un « moyen rhétorique d'ajuster les *claims* et d'anticiper la réponse de l'audience ».

1.3.2 Méthodologies couramment utilisées pour analyser l'*hedging*

Au fil des années, différentes méthodologies ont été proposées pour essayer de détecter et classifier l'*hedging*. Nous regroupons les travaux les plus importants ci-dessous. Ils nous ont inspiré·es et nous ont permis d'opter pour l'apprentissage non-supervisé en toute connaissance de cause. De plus, la liste d'indices qui nous a permis de détecter les *claims* et la modalité épistémique a été créée à partir des différentes listes proposées dans les études précédemment menées que nous présentons dans cette sous-section.

Kilicoglu and Bergler [2008] focalisent ainsi leur travail sur les articles de recherche biomédicale et le langage spéculatif qui y est employé. Leur approche est néanmoins beaucoup plus motivée par la linguistique et ils font notamment appel à la notion de modalité épistémique. Ils centrent leur étude sur l'utilisation des verbes modaux, et utilisent `WordNet` et `UMLS SPECIALIST Lexicon` pour augmenter leur liste d'indices d'*hedging*, qu'ils avaient établie à partir de celle de Hyland [1998] et enrichie avec leurs observations personnelles. Ils établissent ensuite cinq forces d'*hedging*, qui reposent notamment sur des motifs syntaxiques. Leur conclusion est la suivante : « les stratégies d'*hedging* utilisées dans les articles scientifiques sont basiques et prévisibles, elles servent à adoucir les *claims* ou indiquer l'incertitude et elles peuvent être capturées en utilisant une combinaison de moyens lexicaux et syntaxiques » [9].

Ce bilan confirme une fois de plus que l'utilisation d'une liste d'indices centrée sur les verbes modaux permet une extraction satisfaisante des *hedges* et nous incite donc à

9. *« Our results confirm that writers of scientific articles employ basic, predictable hedging strategies to soften their claims or to indicate uncertainty and demonstrate that these strategies can be captured using a combination of lexical and syntactic means »*

utiliser cette approche. C'est également le cas des études suivantes, qui construisent leurs listes d'indices de plusieurs façons :

Dans les corpus de Hyland [1998], les indices les plus présents sont : *indicate, would, may, suggest, could, about, appear, might, likely, propose, probably, apparently, should, seem, possible* et 42 % des instances de *hedges* du corpus sont présentes dans la même phrase qu'au moins un autre indicateur. Nous ajoutons ces indices à notre liste.

Mercer et al. [2004] proposent également une liste d'*hedges* : verbes et adverbes modaux, verbes lexicaux épistémiques, quantifieurs indéfinis [...] ; « tout ce qui contribue à créer un contexte rhétorique et interpersonnel qui cherche à anticiper le rejet du lectorat » [10]. Après observation de données, ils concluent que le phénomène est distribué de manière inégale selon les différentes sections des articles scientifiques, et catégorisent également les *hedges* en deux selon la terminologie de Hyland [1998] (*« content-oriented hedges » et « reader-oriented hedges »*).

Conway et al. [2009] mettent en place des classifieurs automatiques binaires qui catégorisent une phrase comme spéculative ou non. Ils améliorent leurs systèmes en utilisant une liste de 105 lexèmes indicateurs d'*hedging* (empruntée à Mercer et al. [2004]) et l'utilisent pour créer une métrique de spéculation qui, à partir des fréquences des indices d'*hedging*, estime à quel point un article de presse est spéculatif.

L'étude de Ganter and Strube [2009] détecte l'*hedging* grâce à des tags Wikipedia, aux phrases ambiguës (« *weasel phrases* ») et à des caractéristiques linguistiques telles que certains adverbes, les constructions passives et les sujets numériquement non spécifiés.

La catégorisation de l'*hedging* proposée par Light et al. [2004] nous intéresse. Leur étude porte sur les spéculations dans la bioscience. Tout comme leurs prédécesseurs, ils soulèvent des problèmes liés à l'annotation manuelle. Toutefois, leur travail nous semble particulièrement pertinent car il se base sur trois catégories, qui sont proches de celles que nous finissons par mettre en place dans notre système : faiblement spéculatif, hautement spéculatif, et définitifs. Ils estiment ainsi que dans leur corpus, 82 % des phrases sont définitives et seulement 18 % spéculatives.

Finalement, des méthodes différentes de celles utilisées jusqu'à présent sont présentées dans les derniers articles de cette Section :

La *shared task* de CoNLL-2010 portait sur la détection de l'incertitude (à travers l'*hedging*) dans les articles scientifiques biomédicaux et est présentée dans [Farkas et al., 2010]. Au total, parmi les 23 participations, on compte « six systèmes utilisant une classification classique par phrases avec sacs de mots ; les autres équipes se sont concentrées sur les phrases de repérage [...], classifiant une phrase comme incertaine si elle contenait au moins une phrase de repérage » [11].

10. *« all contriving to create a rhetorical and interpersonal context which seeks to pre-empt the reader's rejection »*

11. *«[...] 6 systems with classical sentence classification with bags of words. The remaining teams focused on the cue phrases and sought to classify every token if it was a part of a cue phrase, then a*

Kilicoglu and Bergler [2010] présentent leur participation à cette tâche. Ils ont utilisé une approche à base de règles qui repose à la fois sur des informations lexicales (contenues dans un simple dictionnaire) et syntaxiques (extraites à partir du `Stanford Lexicalised Parser`).

Medlock and Briscoe [2007] proposent un système d'apprentissage faiblement supervisé pour la classification d'*hedges* dans la littérature scientifique. Après un travail minutieux de définitions et la création d'un guide d'annotation, ils mettent en place une classification supervisée binaire de phrases, qui repose notamment la présence d'indices d'*hedging*. Dans leur *baseline*, basée sur une technique de *substring matching*, une phrase est catégorisée comme spéculative si elle contient au moins un des indices de leur liste. Ce travail se distingue des autres car il est faiblement supervisé. Nous pouvons donc nous en inspirer plus largement, et la méthodologie que nous adoptons finalement repose en effet sur du *substring matching* réalisé à partir d'une liste d'indices.

Finalement, l'étude la plus récente que nous avons trouvée sur ce sujet est celle de Szarvas et al. [2012]. Ils présentent un nouveau système de classification faiblement supervisé d'*hedges* dans des textes biomédicaux, basé sur des SVM ainsi que des bigrammes et trigrammes de mots. Selon eux, nous pouvons conclure des travaux précédents sur le sujet que « la détection de l'*hedging* peut être résolue efficacement en recherchant des mots-clés spécifiques qui impliquent que le contenu d'une phrase est spéculatif et en construisant des règles expertes simples qui décrivent les circonstances dans lesquelles un mot-clé doit apparaître et comment il doit apparaître » [12].

1.3.3 Enjeux soulevés

Deux des études précédemment mentionnées abordent les enjeux soulevés par l'*hedging* dans les écrits scientifiques et offrent une réflexion sur la science dans sa généralité.

Hyland [1998] met ainsi l'accent sur les raisons inter-personnelles et institutionnelles qui poussent les auteur·ices à utiliser ce procédé.

Il rejette la croyance commune selon laquelle les articles scientifiques seraient rédigés dans le seul but de partager une vérité de manière purement objective, impersonnelle et informative. Au contraire, il affirme que ce type d'écrits recèle de nombreux enjeux rhétoriques. Les auteur·ices essayent en effet de persuader de la justesse et de la véracité de leurs affirmations afin de s'assurer que celles-ci soient acceptées et ratifiées comme nouvelles connaissances dans la communauté disciplinaire, mais également qu'elles leur octroient de la reconnaissance et du crédit. La recherche n'est alors plus une quête de la vérité, mais une quête du consensus : il faut que le lectorat soit d'accord avec les interprétations proposées.

Les *hedges* participent grandement à accomplir ces missions, ils remplissent en effet « une fonction de persuasion en proposant une interprétation plausible basée sur les données expérimentales afin d'entamer le processus qui transforme l'interprétation subjective

sentence was predicted as uncertain if it contained at least one recognized cue phrase.»)

12. *« Previous studies showed that the detection of hedging can be solved effectively by looking for specific keywords which imply that the content of a sentence is speculative and constructing simple expert rules that describe the circumstances of where and how a keyword should appear. »*

en fait scientifique » [13]. Les *hedges* permettent ainsi de présenter les affirmations avec précaution, précision et humilité.

Hyland adopte ensuite une approche pragmatique et estime qu'il y a une part d'inconscient dans l'*hedging*. En effet, on ne sait pas toujours identifier la motivation précise pour laquelle on le réalise. Il dresse également une classification des *hedges* : ceux qui sont orientés vers le contenu (*« content-oriented hedges »*) et ceux qui sont orientés vers le lectorat (*« reader-oriented hedges »*). Il admet cependant qu'une catégorisation précise des *hedges* scientifiques est impossible, car la subjectivité est trop importante. Cette remarque coïncide avec la difficulté d'annotation manuelle des *claims* que nous avons pu expérimenter et qui était rapportée dans des articles de la Section précédente (voir 1.1).

En conclusion, l'auteur suggère que l'*hedging* est utilisé « inter-personnellement, pour reconnaître le droit du lecteur à réfuter les affirmations, et ce en marquant les déclarations comme provisoires jusqu'à ce qu'elles soient acceptées par les pairs, mais aussi institutionnellement, en permettant de démontrer un engagement envers les normes d'une fraternité de scientifiques dans laquelle la déférence, le débat, une approche partagée des valeurs de vérité et le respect des opinions d'autrui sont des principes de communication appréciés » [14]. Il souhaite ainsi mettre la lumière sur les aspects positifs de l'*hedging*.

Cette thèse est utile à notre étude à différents niveaux. Tout d'abord, elle fournit des définitions claires des concepts que nous utilisons ainsi que des exemples concrets d'*hedges*. Par la suite, l'auteur développe tout un argumentaire sur les raisons d'utilisation d'*hedges* dans la recherche scientifique. Cela nous permet d'obtenir une vision encore plus fine du concept ainsi que des enjeux qu'il recouvre, qui dépassent le simple désir de reconnaissance, s'inscrivent dans toute un système institutionnel et permettent une réflexion sur le concept même de science.

De la même manière, Mercer et al. [2004] se focalisent sur l'*hedging* dans les citations des écrits scientifiques en mettant en lumière ses enjeux.

Selon eux, la recherche scientifique repose essentiellement sur l'acceptation et l'intégration des nouveaux résultats par la communauté. Les membres de cette communauté doivent alors se persuader mutuellement de la validité de leurs résultats. Ainsi, « l'utilisation de l'*hedging* dans l'écriture scientifique fait en réalité partie d'un objectif pragmatique plus large de la part de l'auteur : il doit à la fois avancer des revendications qui doivent être considérées comme dignes d'être publiées [...], et faire attention à présenter son travail comme acceptable pour sa communauté sociale de pairs universitaires et constituant une continuation de la connaissance établie » [15].

13. *« Hedges serve the persuasive function of proposing a plausible interpretation based on experimental data in order to begin the precess which transforms subj interpretation into scientific fact »*

14. *« Interpersonally, hedges are employed to acknowledge the reader's right to refute claims by marking sattements as provisional until accepted by peers. Institutionally, they assist writers in creating an appropriate professional persona, demonstrating a commitment to the norms of a fraternity of scientists in which deference, debate, a shared approach to truth values, and respect for the views of others are valued communicative principles. »*

15. *« The use of hedging in scientic writing is actually part of a larger pragmatic purpose on the part of the author : she is simultaneously putting forth claims that must be seen as worthy of publication or as a basis for funding, while at the same time she must be careful to present her work as acceptable to her social community of academic peers and as constituting a continuation of established knowledge »*

Chapitre 2

Méthodologie

Notre travail repose sur plusieurs grandes étapes que nous présentons dans ce chapitre.

Tout d'abord, nous construisons notre corpus et lui appliquons différents pré-traitements (voir Section 2.1). Ensuite, nous effectuons l'extraction des *claims* ainsi que le *clustering* (voir Section 2.2 et 2.3). Finalement, pour tester d'autres hypothèses sur les corrélations avec des caractéristiques propres aux auteur·ices, nous procédons à l'extraction de leur genre, de leur continent d'origine et de leurs éventuelles affiliations à des institutions de prestige (voir Section 2.4). Les sous-corpus ainsi créés seront ensuite recoupés avec les résultats des *clusterings*. Nous récapitulons les différentes étapes sur la Figure 2.1.

2.1 Présentation du corpus, des outils et des pré-traitements

Nous illustrons toutes les étapes décrites dans cette section grâce à un schéma (voir Figure 2.2).

Nous avons construit un corpus de 6 372 articles en anglais, publiés entre 1979 et 2020 dans le cadre de la conférence de TAL internationale ACL. Nous sélectionnons uniquement les articles courts et longs de la conférence principale. Cela représente 216 689 333 de tokens.

Nous avons pour cela récupéré les fichiers TXT des articles de 1979 à 2016 par l'*AAN Anthology Network Corpus* [Radev et al., 2013] et avons converti nous-mêmes les fichiers PDF des articles postérieurs grâce au script de Dallas Card [1].

Nous avons ensuite procédé au découpage des articles selon la structure traditionnelle des articles scientifiques : résumé (*abstract*), introduction, corps, conclusion. La partie corps inclut alors tout ce qui est compris entre l'introduction et la conclusion. Nous ne l'affinons pas car notre première idée était d'inspecter les *claims* des autres parties. De plus, les corps d'articles ont des structures très hétérogènes. Détecter les sous-parties demanderaient un travail supplémentaire qui n'est ici pas au coeur de notre étude, nous ne n'y attardons donc pas et gardons un seul bloc de corps.

Puis nous avons effectué le découpage en phrases en utilisant de simples *split* selon les points. Nous utilisons ensuite une fonction de pré-traitement qui enlève les liens hypertextes, les caractères spéciaux, les nombres, les mots grammaticaux (d'après la liste de mots outils pour l'anglais de NLTK) et met tout en minuscules. Notons toutefois que le filtrage des mots grammaticaux est optionnel, et qu'il n'est pas toujours réalisé lors de nos tests afin de mesurer sa pertinence.

Dans un premier temps, nous nous sommes tenus à ces pré-traitements, puis, afin d'essayer d'améliorer les résultats, nous avons décidé de lemmatiser notre corpus. Nous

1. www.github.com/dallascard/acl-papers

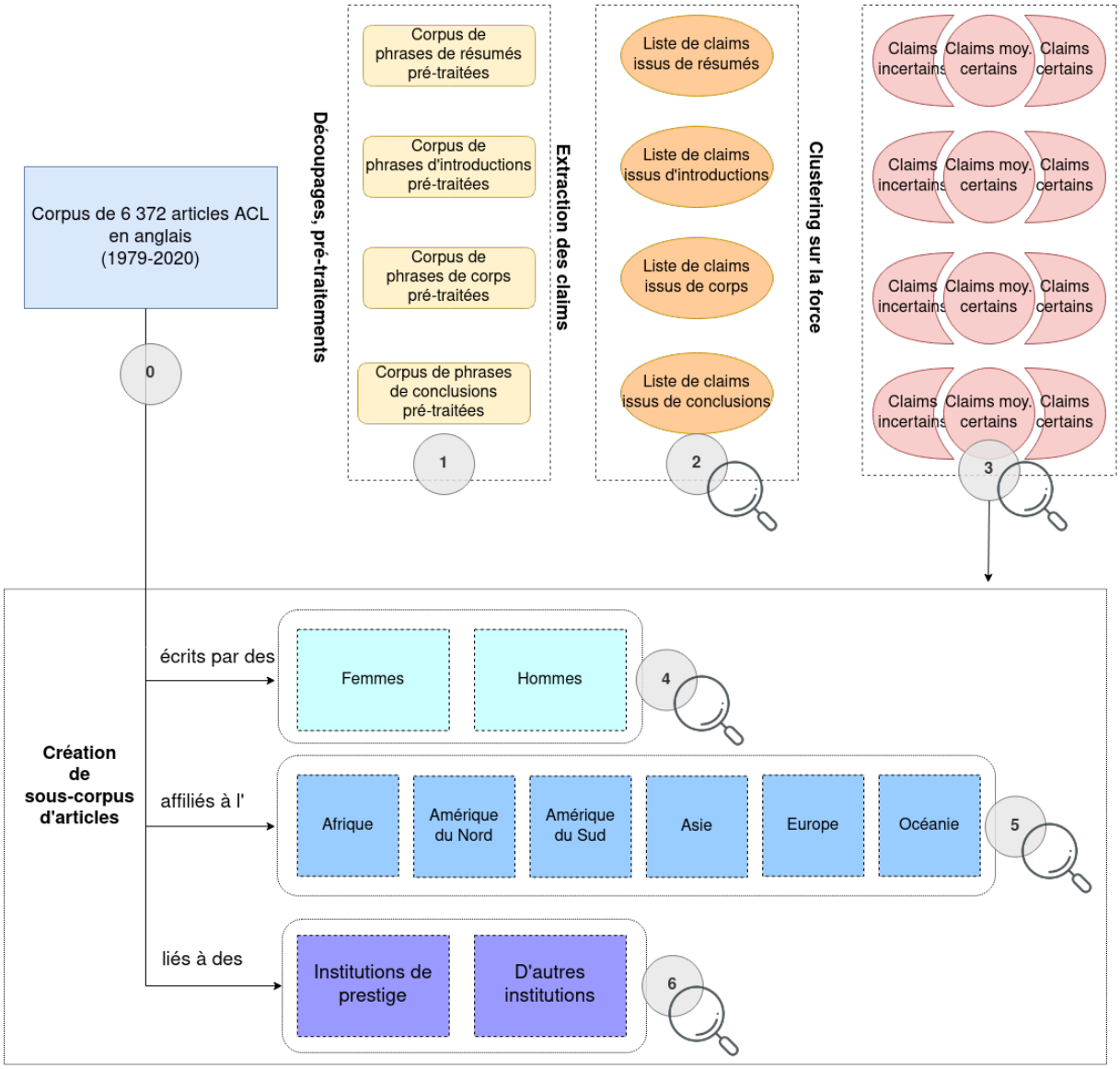

FIGURE 2.1 – Schéma récapitulatif des différentes étapes du travail

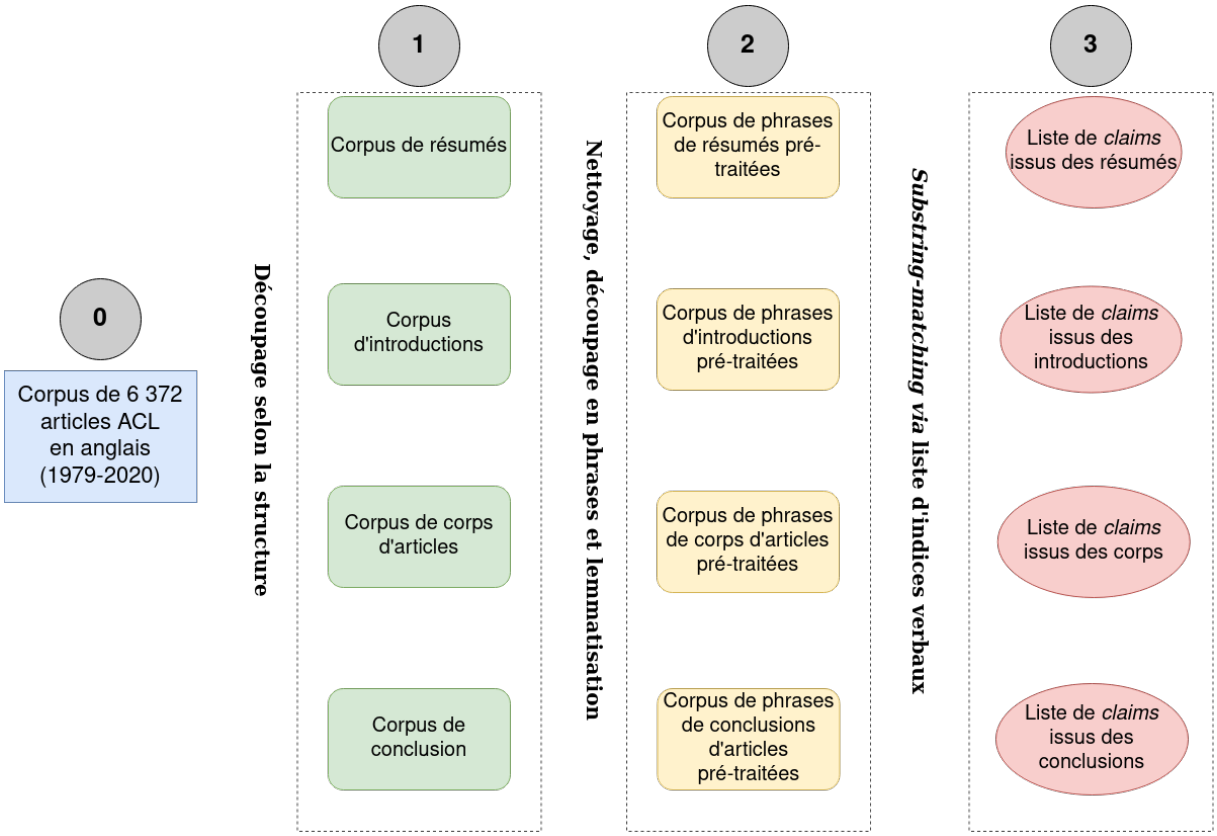

FIGURE 2.2 – Etapes de pré-traitement d'extraction des *claims*

avons d'abord essayé de lemmatiser avec `WordNetLemmatizer` de `NLTK` [2], mais les résultats ne semblaient pas satisfaisants. En effet, après vérification manuelle, nous avons détecté qu'un grand nombres de tokens n'avaient pas été lemmatisés. Nous avons donc finalement opté pour le lemmatiseur de `StanfordCoreNLP` [3].

Cette lemmatisation s'est avérée plus efficace, mais également plus chronophage et gourmande en ressources. Il a donc fallu la réaliser en utilisant le serveur fourni par Grid'5000 [4], auquel j'ai pu avoir accès grâce à mon affiliation au LORIA.

Deux autres outils ont également été cruciaux dans notre étude. D'une part, la bibliothèque libre Python `scikit-learn`, spécialisée dans l'apprentissage automatique, qu'il soit supervisé ou non. D'autre part, le logiciel de textométrie `TXM` [5], qui propose de multiples fonctionnalités d'analyses de corpus : construction de partitions et de sous-corpus, calcul des fréquences des différents termes du vocabulaire et de la progression d'une unité dans un corpus, production de spécificités statistiques, ... Dans notre travail, nous utiliserons uniquement la fonction de spécificités, c'est-à-dire de calculs des mots particulièrement présents dans une partition, et donc potentiellement significative et représentative de cette partie du corpus.

Nous utilisons et adaptons également des scripts Python pré-existants :

— le tutoriel d'Andrew D sur le *clustering* avec Python et TF-IDF [6]

2. https://www.nltk.org/api/nltk.stem.wordnet.html
3. https://stanfordnlp.github.io/CoreNLP/
4. https://www.grid5000.fr/w/Grid5000:Home
5. https://txm.gitpages.huma-num.fr/textometrie/
6. https://medium.com/mlearning-ai/text-clustering-with-tf-idf-in-python-c94cd26a31e7

— le tutoriel de B. Bonaros sur la méthode Elbow [7]

2.2 Extraction des *claims*

La première étape véritablement propre à notre étude est l'extraction de *claims*. Nous avons commencé par extraire les *claims* présents seulement dans les résumés des articles avant d'étendre à toutes les autres parties (citées plus haut).

D'une manière semblable à celles utilisées par la plupart des études dans notre état de l'art (voir Sections 1.1, 1.3 et en particulier la citation de Szarvas et al. [2012]), notre approche est basée sur du *substring matching* et repose ainsi sur une liste d'indices d'*hedging*. Il s'agit en effet de la méthode qui fonctionne le mieux, et, bien qu'une telle méthode basée sur une liste demeure arbitraire, nous avons essayé de la rendre la plus pertinente possible. Nous utilisons la méthode de *substring matching* intégrée à Python 3, qui est basée sur l'algorithme de Boyer-Moore-Horspool [8].

Il a donc fallu nous mettre d'accord sur une liste de tokens indiquant un *claim* afin de pouvoir extraire les phrases contenant au moins l'un de ces tokens. Au début de notre travail, nous avons établi une liste assez restreinte, basée sur des observations menées suite à la détection manuelle de *claims* dans quelques articles de TAL :

can, could, show, prove, demonstrate, find, lead, significant.

Nous avons par la suite souhaité retravaillé et étoffé cette liste. Nous avons essayé plusieurs approches : nous avons extrait les synonymes de la liste précédente grâce à `WordNet` [9] d'une part, et `Gensim` [10] de l'autre, nous avons utilisé des méthodes de *Web-Scraping* sur le site `WordReference`, et nous nous sommes intéressés à la polarité des phrases.

Toutefois, ces méthodes ne nous ont pas semblé satisfaisantes. En effet, les listes ainsi générées étaient très longues et comportaient beaucoup de bruit, car ce sont des verbes très communs, qui ont beaucoup de synonymes alors que ce sont en réalité des acceptions précises de nos verbes de base que nous souhaitons prendre en compte.

Nous avons fini par établir une liste à partir des indices cités dans les articles que nous avons lus pour nos états de l'art sur la modalité épistémique et l'*hedging*. Cela nous semble être une décision pertinente, fondée sur d'autres travaux, et donc moins arbitraire qu'une liste construite uniquement à partir de nos intuitions.

Nous avons ensuite choisi de ne garder que les verbes de cette nouvelle grande liste car ils constituent la partie du discours la plus représentative de la modalité épistémique et des *claims*. En effet, si l'on tente d'extraire des *claims* avec des indices de toutes les parties du discours, on se rend compte qu'en réalité un peu plus de 75 % d'entre eux ont été détectés grâce à des indices verbaux. Ce nombre entre d'ailleurs en accord avec les remarques de Hyland [1998] (p.104), qui calcule « les fréquences relatives des catégories grammaticales utilisées pour la modalité épistémique » et trouve que les verbes modaux constituent 40,2 % de son corpus, les verbes lexicaux 33,3 % (soit 73,5 % d'indices verbaux), puis 18,1 % d'adverbes, 4 % d'adjectifs et 4,5 % de noms. Ces données sont également proches

7. https://predictivehacks.com/k-means-elbow-method-code-for-python/
8. https://stackoverflow.com/questions/18139660/python-string-in-operator-implementation-algorithm
9. https://wordnet.princeton.edu/
10. https://radimrehurek.com/gensim/

Verbes	*find, can, could, may, might, must, should, claim, admit, discover, suggest, predict, prove, show, explain, infer, conclude, demonstrate, lead, succeed, intend, indicate, assume, appear, seem, favor, think, believe, analyze, examine, report, dare, point out, note, assert, declare, remark, comment, observe, reveal, disclose, confirm, convince*
Adverbes	*certainly, necessarily, apparently, probably, likely, possibly, presumably, seemingly, firmly, perhaps, obviously, definitely, indeed, presumably, surely, undoubtedly, evidently, significantly, remarkably, admittedly, assuredly, incontestably, indisputably, indubitably, unarguably, undeniably, undoubtedly, unquestionably, clearly, greatly, manifestly*
Adjectifs	*certain, necessary, sure, apparent, probable, presumed, hypothetic, significant, able, evident, clear, possible, potential, amazingly, greatly, perfectly, unbelievably, crucial, best, most, unprecedented*
Noms	*evidence, proof, guarantee, proposal, likelihood, allegation*

TABLEAU 2.1 – Listes d'indices de modalité épistémique en anglais, utilisées pour extraire les *claims*

de celles données par Holmes [1982], qui trouve que la certitude est exprimée à 37 % par des auxiliaires modaux, 36 % par des verbes lexicaux (soit 73 % d'indices verbaux également), 12 % par des constructions adverbiales, 8 % par des noms et 7 % par des adjectifs.

De plus, nous nous rendons compte que les supposés *claims* détectés avec des indices non verbaux ne sont en réalité pas des *claims*. Il semblerait donc que les indices non verbaux entraînent plus de bruit que de véritables *claims*, c'est pourquoi nous avons décidé de ne pas les inclure.

Enfin, nous procédons à l'extraction des *claims* grâce à du *pattern-matching* à partir de cette nouvelle liste d'indices. Nous en profitons également pour regarder les indices les plus populaires ainsi que les co-occurrences (que nous présenterons dans le chapitre Résultats, 3.1).

2.3 *Clustering*

Afin de classifier les *claims* précédemment extraits, nous avons choisi d'utiliser des techniques d'apprentissage non-supervisé, et plus précisément de *clustering*.

L'apprentissage supervisé s'est révélé très difficile à mettre en place car il requiert un processus d'annotation manuelle. Or, l'annotation manuelle des phénomènes de modalité épistémique et des degrés de certitude dans les textes est une tâche compliquée et chronophage. Lors de son étude, Rubin [2006] indique que « la charge de travail liée à l'annotation a été estimée à 10 heures [par personne, pour l'annotation de 10 articles] ». De plus, les scores d'accord inter-annotateurs sont relativement peu élevés. Par exemple, toujours chez Rubin [2006], les expériences se soldent toujours par un Kappa de Cohen [Cohen, 1960] oscillant entre 0,13 et 0,65 mais dépassant en réalité rarement les 0,5 et le « pourcentage d'accord observé » maximal s'élève à 71 %. Nos tentatives ont confirmé cette difficulté. Nous choisissons alors de nous consacrer pleinement à l'apprentissage non-supervisé.

L'idée de créer une *baseline* en classant nous-mêmes les différents indices verbaux a également été vite abandonnée pour la même raison : il s'avère difficile de décider quel verbe reflète quel degré de certitude.

Histoire et théorie du *clustering*

Une grande partie de notre travail réside ainsi dans la mise en place d'un système d'apprentissage complètement non-supervisé grâce à du *clustering*. Nous entrons un peu plus dans les détails théoriques de cette méthode dans cette sous-section.

Dans le chapitre 14 de leur ouvrage, Manning and Schutze [1999] présentent les notions de *clustering* hiérarchique et non-hiérarchique. Ils estiment que l'approche non-hiérarchique est préférable si les jeux de données sont conséquents, et que *K-means* est la méthode la plus simple conceptuellement et devrait être utilisée en premier. C'est celle-ci que nous utiliserons lors de notre travail. Cet algorithme de *clustering* suppose un simple espace de représentation euclidien et définit les *clusters* par le centre de masse de leurs membres. La méthode est détaillée : des centres de *clusters* sont définis, puis, chaque objet se voit affecté au *cluster* dont le centre est le plus proche. Chaque centre est ensuite recalculé en fonction de ses membres. Selon eux, cette méthode permet de mettre en évidence les types de mots et les relations présentes dans les données.

Eissen and Stein [2002] décrivent également le principe du *clustering* : « Il tente d'identifier des groupes au sein d'un ensemble d'objets, de telle sorte que les éléments des différents groupes présentent des différences significatives en ce qui concerne leurs caractéristiques métriques. Les algorithmes de *clustering* fonctionnent sur les similarités entre objets, qui sont à leur tour calculées à partir de descriptions abstraites des objets. Chacune de ces descriptions est un vecteur d de nombres comprenant les valeurs des caractéristiques essentielles des objets » [11].

Ils abordent aussi les métriques d'évaluation de l'index de Dunn et de la méthode Elbow. Ils concluent que les algorithmes de *clustering* permettent de relever les deux défis posés par la classification automatique de texte que sont la formation efficace de catégories et l'absence de schéma de catégorisation préalable.

Les définitions données par Schaeffer [2007] sont également pertinentes. Elle estime que « le but du *clustering* est de diviser l'ensemble de données en *clusters*, de sorte que les éléments assignés à un *cluster* particulier soient similaires ou connectés dans un sens prédéfini » [12].

Récemment, M. Salih and Jacksi [2020] dressent un état de l'art des algorithmes de *clustering* de documents basés sur la similarité sémantique. On peut en retirer que *K-means* est autant utilisé que le *clustering* hiérarchique, mais que *Fuzzy C-means* et *bisecting K-means* sont moins populaires. L'utilisation de `Wordnet` est également grandissante.

11. *« It tries to identify groups within an object set such that elements of different groups show significant differences with respect to their metric features. Clustering algorithms operate on object similarities, which, in turn, are computed from abstract descriptions of the objects. Each such description is a vector d of numbers comprising values of essential object features. »*

12. *« Formally, given a data set, the goal of clustering is to divide the set into clusters such that the elements assigned to a particular cluster are similar or connected in some predefined sense. »*

Ces remarques nous permettent d'opter pour *K-means* et nous donnent l'idée de mener des expériences avec `Wordnet`.

Implémentation et évaluation

Ainsi, nous avons commencé à mettre en place du *clustering*. Nous avons fait des essais avec deux librairies : `Gensim` et `sklearn`. Nous avons finalement retenu cette dernière et poursuivi nos expériences, notamment en faisant varier le nombre de *clusters* entre deux et dix.

Pour entrer plus amplement dans les détails, nous avons utilisé le vectoriseur `TF-IDF`, la méthode de partitionnement en k-moyennes (`K-means`) et l'analyse en composantes principales (*Principal Component Analysis -PCA-*). Nous avons également fait en sorte de récupérer les mots-clefs représentatifs de chaque *cluster* et de visualiser les clusters (grâce à la librairie `seaborn`). Notre code est en grande partie inspiré et adapté du tutoriel d'Andrew D sur medium.com [13].

La question de l'évaluation des techniques d'apprentissage non-supervisé s'est ensuite posée. En effet, l'évaluation traditionnelle en apprentissage supervisé repose largement sur l'adéquation avec les données de référence, c'est-à-dire les données annotées manuellement, dont nous ne disposons pas ici.

Après avoir fait quelques recherches, nous cherchons à calculer les scores Silhouette avec différentes métriques (*cityblock, cosinus, euclidienne, l1, l2, Manhattan*) ainsi que les scores Calinski-Harabasz et Davies-Bouldin. Ces trois métriques sont calculées en utilisant la librairie `sklearn.metrics` et les distances à l'aide de `scipy.spatial.distance`.

Ensuite, nous avons analysé manuellement les mots-clefs afin de relier chacun des *clusters* à un degré de certitude et avons réassigné chaque numéro de cluster au degré correspondant (en partant du principe que zéro est le moins certain et que le degré de certitude augmente à chaque numéro de cluster) pour homogénéiser les résultats obtenus dans les différentes parties des articles afin de pouvoir effectuer des comparaisons.

Nous réalisons par la suite des calculs et des graphiques afin de pouvoir analyser et comparer les *claims* selon la force qui leur a été attribuée et la partie de l'article où ils se situent. De plus, nous utilisons `TXM` [14] afin de regarder les spécificités de chaque sous-corpus de *claims* selon la catégorie que le *clustering* leur a assignée. Cela nous permet de mener une analyse linguistiquement plus fine de nos *claims*, mais aussi de valider les *clusters* obtenus avec nos scripts.

2.4 Préparation des autres approches

Nous développons des méthodes pour pouvoir prendre en compte d'autres facteurs dans nos analyses et recoupons ensuite les informations récoltées sur les auteur·ices avec les résultats du *clustering* pour pouvoir effectuer des comparaisons.

13. https://medium.com/mlearning-ai/text-clustering-with-tf-idf-in-python-c94cd26a31e7
14. https://txm.gitpages.huma-num.fr/textometrie/

2.4.1 Corrélation avec le genre des auteur·ices

La sous-section suivante traite du concept social de genre. Nous sommes conscient·es que notre approche est quelque peu simpliste, notamment du fait qu'elle ne prend en compte que les deux genres hégémoniques et qu'elle part du principe qu'un prénom égale un genre. Toutefois, nous pensons qu'elle demeure pertinente dans la mesure où les stéréotypes et discriminations de genre s'appliquent généralement aux personnes ayant un prénom traditionnellement perçu comme féminin, qui ont très probablement été assignées femmes à la naissance ou qui utilisent un prénom féminin à dessein, afin qu'il colle mieux à leur identité de genre. De plus, nous essayons de suivre les recommandations de Larson [2017] au mieux.

Plusieurs études suggèrent que les femmes seraient moins assertives, auraient moins confiance en elles et feraient preuve de plus de politesse [Lakoff, 1973b] ; elles pourraient avoir donc plus tendance à avoir recours à des parades linguistiques liées à la modalité épistémique comme l'indique J. Coates dans *The role of epistemic modality in women's talk* [Facchinetti et al., 2012]. Nous avons souhaité regarder si tel était le cas dans notre corpus et notre classification.

Pour cela, nous avons dû trouver un moyen de détecter le genre des auteur·ices automatiquement. Nous avons essayé d'utiliser le module de détection d'entités nommées de `Spacy`, mais il s'est avéré peu performant dans le sens où peu de prénoms de notre corpus étaient reconnus. Nous avons donc fusionné plusieurs listes [15] associant un prénom à un genre afin d'avoir des données représentatives d'un maximum de régions du monde. Nous arrivons à un total de quasiment 400 000 prénoms assez équitablement répartis en prénoms typiquement masculins d'une part et féminins de l'autre.

Nous avons alors utilisé des fichiers BIB convertis en CSV afin d'extraire les listes d'auteur·ices des différents articles.

Puis, nous avons comparé ces listes d'auteur·ices aux prénoms présents dans notre liste afin de les associer au genre et à l'article correspondant. Au total, nous avons repéré 11 132 prénoms masculins et 3 066 féminins, mais comme les articles sont souvent rédigés par plus d'une personne, cela couvre en fait 4 432 articles du corpus, soit un peu plus de 69,5 %.

Nous avons ensuite relié ces données aux résultats du *clustering* afin de pouvoir mettre en correspondance le degré de certitude utilisé dans tel article et le genre des auteur·ices de l'article en question.

Nous avons ainsi pu effectuer des comparaisons selon deux approches : soit en prenant en compte le genre du premier ou de la première auteur·ice, soit en prenant en compte le genre le plus représenté dans le groupe d'auteur·ices. Par exemple, un article écrit par ce que nous détectons comme deux femmes et un homme sera comptabilisé dans la catégorie femme. Cette approche prévoit alors une catégorie égalité pour les cas où autant d'hommes que de femmes font partie des auteur·ices.

15. A noter : des vérifications et corrections manuelles ont été effectuées, notamment pour les prénoms chinois. Listes trouvées sur https://raw.githubusercontent.com/ellisbrown/name2gender/master/data/name_gender_data.csv, et https://raw.githubusercontent.com/hadley/data-baby-names/master/baby-names.csv

2.4.2 Corrélation avec le continent

Nous nous sommes également intéressé·es à la corrélation qui pourrait exister avec le continent d'affiliation des auteur·ices, en prenant en compte celui qui est le plus présent parmi le groupe d'auteur·ices.

Nous adoptons une méthodologie semblable à celle utilisée précédemment pour le genre, en effectuant du *pattern-matching* dans les 1000 premiers caractères des articles. Ainsi, si un nom de pays y est détecté, nous le comptabilisons et considérons qu'il fait partie des pays affiliés aux auteur·ices de l'article.

Pour cela, nous nous basons sur un fichier JSON [16] contenant un dictionnaire qui, à chaque pays listé, associe différentes informations, dont le continent auquel il appartient.

Toutefois, après avoir utilisé cette méthode, nous avons constaté que le nombre d'articles originaires des États-Unis semblait étrangement bas. Nous avons alors remarqué que les auteur·ices états-unien·nes omettent souvent de mentionner clairement leur pays d'affiliation. En revanche, on peut déduire qu'il s'agit bien des États-Unis car c'est leur état ou leur université qui est mentionnée, ou bien le nom de domaine *.edu* à la fin de leur adresse électronique. Nous ajoutons donc une liste contenant des noms d'universités et d'états très populaires, qui, s'ils sont détectés, associent l'article aux États-Unis :

.edu, Google, Facebook, Microsoft, California, Washington, Stanford, Harvard, Amazon, Pennsylvania, Brown University, New York.

2.4.3 Corrélation avec des affiliations à des institutions de prestige

Enfin, nous avons réutilisé la technique de *pattern-matching* sur les 1000 premiers caractères des articles pour extraire des grandes institutions, à savoir ici les GAFAM (Google, Amazon, Facebook, Apple, Microsoft) et les 50 premières universités du classement de Shanghai de 2021 [17]. La liste complète est donnée en Annexes (voir Chapitre 6).

Nous avons décidé d'inclure les GAFAM car le TAL joue un rôle de plus en plus central au sein de ces entreprises, et elles financent alors un grand nombre de conférences et de laboratoires du domaine. Toutefois, comme le suggèrent Abdalla and Abdalla [2021], la présence des BigTech (dont font partie les GAFAM) dans la recherche n'est pas sans conséquence.

D'autre part, nous ajoutons à notre liste les universités reconnues par le classement de Shanghai car elles sont mondialement réputées, et cette popularité pourrait avoir un impact sur la manière dont les personnes qui y sont affiliées rédigent leurs articles.

16. Trouvé sur https://github.com/annexare/Countries/blob/master/data/countries.json
17. https://www.shanghairanking.com/rankings/arwu/2021
Nous pouvons remettre en question le choix de ce classement dont la pertinence scientifique est à débattre. Il demeure toutefois utilisé en France et coïncide avec les universités estimées prestigieuses par le grand public, ce qui le rend utile pour notre approche.

Chapitre 3

Production et analyse des *clusters*

Nous procédons alors à l'extraction des *claims*, puis à l'optimisation du *clustering*. Nous pouvons ensuite commencer l'analyse des *clusters* ainsi obtenus afin de tirer des observations sur les *claims* de notre corpus.

3.1 Tendances de *claims* et co-occurrences

Suite à l'extraction des *claims* telle que détaillée précédemment, nous sommes en mesure de donner quelques chiffres :

Nous trouvons 386 886 *claims* au total dans tout le corpus, dont 26 332 dans les résumés, 33 372 dans les introductions, 279 749 dans les corps et 47 433 dans les conclusions. A noter : nous avons fait le choix de conserver des doublons, c'est-à-dire que si une phrase contient plusieurs indices elle compte plusieurs fois. Cela revient à intensifier le poids de ce *claim* proportionnellement au nombre d'indices qu'il contient. Sans doublons, on a 298 932 *claims* uniques.

Il y a 869 807 indices au total dans les *claims* extraits, qui ont donc été extraits seulement avec des indices verbaux comme expliqué dans la partie Méthodologie 2.2. Il faut également noter qu'un *claim* contient généralement plus d'un indice. Parmi ces indices, 782 887 sont des verbes et 86 920 indices sont non verbaux. Autrement dit, 9,99 % d'indices sont non verbaux et 90,01 % d'indices sont verbaux.

Si l'on s'intéresse aux catégories grammaticales non-verbales :
— 6,76 % des indices totaux et 67,61 % des indices non verbaux sont des adjectifs
— 2,26 % des indices totaux et 22,62 % des indices non verbaux sont des adverbes
— 1,08 % des indices totaux et 10,79 % des indices non verbaux sont des noms

Nous inspectons ensuite les 10 indices verbaux les plus présents dans chaque partie (résumé, introduction, corps, conclusion). Si nous fusionnons ces listes, nous obtenons une liste de 14 verbes, que nous rapportons ci-dessous :
can, show, state, find, may, demonstrate, predict, report, support, lead, should, could, indicate, note

La question des co-occurrences est également intéressante, car comme les chiffres le montrent les *claims* contiennent souvent plusieurs indices. On peut imaginer que c'est notamment le cas des verbes modaux. Nous extrayons donc des tri-grammes, en prenant le token précédent et celui suivant un indice verbal.

Nous filtrons les tokens précédents et suivants les indices verbaux afin de ne garder que les tokens qui font également partie de la liste d'indices (verbaux et non-verbaux inclus, car le but est ici de voir si la modalité épistémique est exprimée de plusieurs façons,

grâce à plusieurs indices). Nous mettons les tableaux correspondant en Annexes (voir Tableaux 6.2).

Avant de clôturer cette section, nous donnons quatre exemples de *claims* extraits ainsi et tirés au hasard afin de donner une meilleure idée du type de phrases sur lesquelles on travaille :

— « Pour être utile, un analyseur syntaxique doit être capable d'accepter une large gamme de types d'entrée, et doit être capable de traiter gracieusement les dysfonctionnements, les faux départs et autres entrées non grammaticales. » [Malouf, 2000] [1]
— « Les résultats expérimentaux montrent que la structure auto-organisée du modèle gram améliore le modèle de base. » [Park et al., 2006] [2]
— « L'expérience montre que RCNN est très efficace pour améliorer l'état de l'art de l'analyse des dépendances sur les jeux de données anglais et chinois. » [Zhu et al., 2015] [3]
— « Les travaux récents sur les modèles tenant compte de la structure ont donné des résultats prometteurs pour la modélisation du langage. » [Zhang and Song, 2019] [4]

Ces phrases illustrent différentes caractéristiques des *claims*. Elles tirent des conclusions à partir de résultats précédemment obtenus ; les auteur·ices estiment alors que leurs systèmes sont « très efficaces » et permettent une réelle « amélior[ation] de l'état de l'art ». Parfois, ce sont même des règles générales voire absolues qui sont formulées, par exemple ici avec le verbe « devoir ».

Ainsi, on y décèle également une portée rhétorique. En effet, ces phrases pourraient nous inciter à privilégier des systèmes et des méthodes plutôt que d'autres. A partir de ces exemples, on pourrait décider d'opter pour le modèle gram, RCNN, ou des « modèles tenant compte de la structure ».

3.2 Optimisation des paramètres de *clustering*

Afin de choisir un paramétrage optimal pour notre *clustering*, nous avons évalué, à l'aide de plusieurs métriques, différentes combinaisons d'outils et de paramètres.

Évaluer la qualité d'un *clustering* est moins intuitif que d'évaluer une méthode supervisée, puisque nous n'avons aucune annotation manuelle, donc aucune référence. Nous nous sommes donc intéressé·es aux métriques adaptées à notre situation.

1. *« In order to be useful, a parser must be able to accept a wide range of input type, and must be able to gracefully deal with dysfluency, false start, and other ungrammatical input. »*
2. *« The experimental result show that the self organize structure of gram model enhance the basic model. »*
3. *« The experiment show that rcnn is very effective to improve the state of the art dependency parse on both english and chinese dataset. »*
4. *« Recent work on structure - aware model have show promising result on language modeling. »*

3.2.1 Déterminer le nombre de *clusters* : méthode *Elbow*

Tout d'abord, l'un des problèmes posés par le *clustering* est la manière de déterminer le nombre de *clusters* optimal. Pour cela, nous utilisons la méthode Elbow (littéralement, méthode du coude), que nous décrivons en réutilisant les travaux d'autres TAListes.

Syakur et al. [2018] semblent être les premiers à présenter cette méthode, qui est à combiner avec *K-means*. Il est expliqué que « la méthode Elbow est facile à mettre en œuvre en regardant le graphique de la valeur idéale de k [nombre de *clusters*] avec la position sur le coude (ainsi que la somme de l'erreur quadratique qui est inférieure à 1). Le meilleur résultat de *clusters* k sera la base du *clustering*. Plus la valeur de la somme de l'erreur quadratique est faible et plus le graphique du coude diminue, meilleurs sont les résultats du *cluster* ». La méthode Elbow porte en effet son nom car elle repose sur un graphique de visualisation qui forme une courbe en forme de bras, et l'on estime que « le nombre de *clusters* optimal est indiqué par l'endroit du graphique qui correspond au coude, là où le graphique a une courbure significative ». ([Saputra et al., 2020])

Shi et al. [2021] détaillent encore le fonctionnement : « on doit généralement effectuer les *K-means* sur le même ensemble de données avec une plage de nombres de *clusters* contigus : [1, L] (L est un nombre entier supérieur à 1). Ensuite, on calcule la somme des erreurs quadratiques (SSE) pour chaque numéro de *cluster* k spécifié par l'utilisateur, en traçant une courbe de la SSE en fonction de chaque numéro de *cluster* k. Enfin, les analystes expérimentés estiment le point de coude optimal en analysant la courbe susmentionnée, c'est-à-dire que le point de coude optimal correspond au numéro de *cluster* optimal potentiel estimé avec une forte probabilité ». Mais l'efficacité de la méthode est nuancée et remise en cause, notamment dans les cas où la courbe est assez régulière et où il est difficile d'identifier ce qui serait le coude. En effet, « le nombre de *clusters* obtenu en utilisant la méthode du coude est un résultat subjectif car il s'agit d'une méthode visuelle, qui ne fournit pas de métrique de mesure pour montrer quel point du coude est explicitement le meilleur. »

Nous implémentons alors cette méthode Elbow avec Python grâce au tutoriel de B. Bonaros[5]. Nous l'avons testé avec plusieurs combinaisons de paramètres mais ne faisons figurer en Annexes (6.3) que les graphiques correspondants au paramétrage finalement utilisé, qui semble indiquer que trois *clusters* seraient pertinents. Nous présentons ici (3.1) le graphique généré à partir du *clustering* sur les conclusions. Sur celui-ci, tout comme sur les autres, nous remarquons en effet que la courbe s'affaisse à partir de k égale trois. Cela correspond donc à notre intuition initiale ainsi qu'à celle de différents auteur·ices de notre état de l'art.

Dès à présent, nous utilisons donc une catégorisation en 3 : 0 étant le degré qui laisse transparaître le moins de certitude et 2 le plus. Autrement dit, le degré 0 serait lié aux *claims* incertains, 1 aux moyennement certains et 2 aux (très) certains.

5. https://predictivehacks.com/k-means-elbow-method-code-for-python/

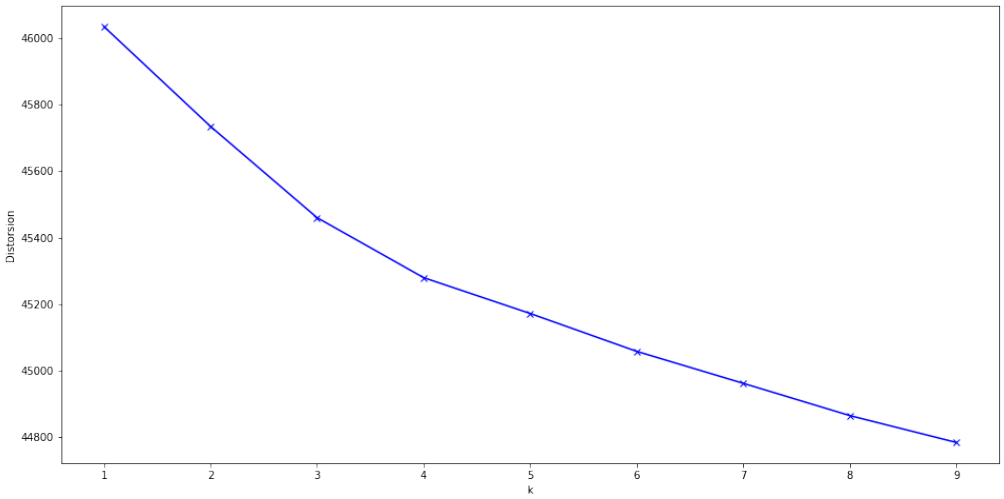

FIGURE 3.1 – Résultat de la méthode Elbow sur les *claims* des conclusions

3.2.2 Déterminer la meilleure méthodologie

Nous utilisons également d'autres métriques d'évaluation, notamment présentées dans [Lallich and Lenca, 2015] et incluses dans le volet *metrics* de `sklearn`[6].

L'indice Calinski-Harabasz [Caliński and Harabasz, 1974] est défini par le ratio de la somme de la dispersion inter-classe et intra-classe. Il favorise les classes compactes et séparables tout en facilitant la détermination du nombre de classes optimal. Il est à maximiser, et compris entre 0 et l'infini.

Le score Davies-Bouldin [Davies and Bouldin, 1979] est « la moyenne des similarités entre chaque *cluster* et le *cluster* le plus similaire » [Lallich and Lenca, 2015]. Ici, il faut tendre vers 0, car les valeurs les plus basses indiquent un meilleur *clustering*, et le nombre de classes n'a pas d'impact sur le calcul.

Enfin, nous utilisons le score Silhouette [Rousseeuw, 1987], compris entre -1 et 1. La silhouette d'un *clustering* est en réalité la moyenne des silhouettes des objets concernés. La silhouette d'un objet i repose sur « la compacité, c'est-à-dire la distance moyenne de l'objet i aux objets de son *cluster*, et la séparabilité, c'est-à-dire le minimum des distances moyenne de l'objet i aux objets de chacune des autres classes ».

Nous calculons les scores Silhouettes avec diverses distances (*cityblock, cosinus, euclidienne, l1, l2, Manhattan*) ainsi que les scores Davies-Bouldin et Calinski-Harabasz, et ce pour chaque partie, à chaque fois sur versions lemmatisée et non lemmatisée, avec et sans mots outils. Nous reportons ces résultats dans des tableaux en Annexes (voir Tableaux 6.4), et mettons ici le tableau des scores sur les conclusions (voir Tableau 3.1). Nous ne faisons toutefois pas apparaître les scores Silhouette avec les métriques *Cityblock, l1* et *l2* pour permettre une meilleure lisibilité. De plus, ces résultats sont toujours très

6. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

	Silhouette Score			CH score	DB score
	Cosinus	Eucl.	Manh.		
Lem., mots outils	**0,01**	**0,005**	-0,02	**260,46**	**10,16**
Lem., pas de mots outils	0,008	0,005	-0,055	246,68	9,05
Pas de lem., mots outils	0,009	0,005	-0,056	240,03	9,86
Pas de lem., pas de mots outils	0,007	0,004	-0,06	224,19	9,47

TABLEAU 3.1 – Résultats des métriques d'évaluation pour le *clustering* sur les conclusions

proches des scores Manhattan (pour Cityblock et l1) et Euclidien (l2).

Nous remarquons que les scores Silhouette demeurent très bas. Cela peut témoigner de la difficulté de la tâche, mais ne doit toutefois pas être décourageant car les visualisations des *clusters* ainsi que les mots-clefs qui leur sont associés sont sémantiquement pertinents.

De plus, les autres résultats des métriques s'avèrent rassurants. En effet, le score de Calinski-Harabasz atteint les 1047 dans les corps d'articles avec lemmatisation et mots outils tandis que le score de Davies-Bouldin descend jusqu'à 8,83 dans la même configuration mais sur les conclusions. Rappelons que pour ce score, il faut tendre le plus possible vers 0 ; plus ce score est bas, meilleur est le *clustering*.

Afin d'avoir un point de comparaison et pour rendre la lecture de ces scores plus aisée, nous avons créé une matrice de même taille mais remplie aléatoirement et avons relancé les calculs. Le score Calinski-Harabasz s'élève à 7,3 seulement, tandis que celui de Davies-Bouldin égale 10,1 sur la matrice aléatoire. Cela prouve bien que notre *clustering* obtient de meilleures performances que le hasard, et est donc pertinent.

Il faut en retenir que les meilleurs scores sont obtenus avec lemmatisation et en conservant les mots outils. C'est donc ce paramétrage que nous utiliserons pour le reste des expériences.

Nous ajoutons brièvement quelques mots et métriques d'évaluation sur la dernière expérience que nous avons menée, lors de laquelle nous avons lancé le *clustering* sur toutes les parties des articles mélangées, donc tous les *claims* d'un coup, en optant pour le paramétrage précédemment mentionné. À l'exception du score Calinski-Harabasz qui s'avère étonnamment élevé, tous les autres scores diminuent. De plus, le temps d'exécution des différents scripts devient extrêmement long, certains n'aboutissent pas après une dizaine d'heures, et supprimer la dimension de comparaison selon les parties des articles réduit beaucoup le potentiel d'analyses et de comparaisons. C'est pour toutes ces raisons que nous décidons de ne pas exploiter plus ce *clustering* et que nous le mentionnons simplement ici.

3.3 Analyse et validation des *clusters*

3.3.1 Validation sémantique des *clusters*

Afin de pouvoir donner une analyse sémantique au *clustering*, nous regardons les 50 mots-clefs considérés comme les plus représentatifs de chaque *cluster*, et ce pour chacun de nos *clusterings* sur chaque partie des articles. Nous reportons ici les 30 premiers mots

(voir Tableaux 3.4), en leur réattribuant manuellement le bon numéro de *cluster*, à savoir : 0 égale *claim* faible, 1 égale moyen et 2 égale fort. Nous avons également réattribué ces numéros dans la suite de nos codes afin que les analyses soient possibles.

Nous mettons en gras les tokens, qui, selon nous, ont un réel poids sémantique dans la catégorisation et auraient pu être attribués ainsi par des personnes lors d'un processus d'annotation manuelle, ou qui font partie de notre liste d'indices. Toutefois, il est logique qu'il y ait peu de mots concernés dans le tableau du degré 0 puisque par définition, c'est l'absence d'indices qui en fait le degré 0.

On remarque cependant que, selon les parties, les *clusters* associés ne sont pas exactement les mêmes, ou du moins que les degrés ne se valent pas totalement, même après notre réattribution manuelle. Ainsi, les résumés et les introductions semblent être les parties où les indices sont les plus nombreux, même au degré 0, tandis qu'ils le sont peu et avec une moindre force pour les conclusions, même au degré 2.

Résumés	*gmail, use, paper, lab, california, on, base, engineering, word, from, with, translation, system, uk, text, machine, laboratory, center, model, cn, school, ac, state, information, china, research, stanford, to,* ***can****, be, language,*
Introductions	*sentence, learn,* ***result****, neural, performance, information, not, text, approach, parse, machine, they, find, or, recent, it, method, this, work, translation, system, art, which, from, language, task, we, with, base, such,*
Corps	*neural, network, datum, dataset, feature, predict, not, sentence, approach, or, performance, such, they, system, language, also, it, method, find, have, result, ha, this, art, task, base, from, word, work, which,*
Conclusions	*acknowledge, technology, fund, contract, also, paper, in,* ***suggestion****, darpa, award, valuable, to, ii,* ***partially****, be, natural, insightful, author, part,* ***would****, nsf, of, like, we, program, helpful, under, project, china, no,*

TABLEAU 3.2 – Top 30 des mots-clefs les plus représentatifs du degré 0 selon les parties

Résumés	*ha, example, structure, but, only, at, task, feature, text, datum, how,* ***predict****, information, base,* ***more****, one, sentence, also, state, system, have, such, these, or,* ***may****,* ***find****, language, word, they, on,*
Introductions	*structure, if, however, only, also, ha, we, user,* ***could, must,*** *at, but, other,* ***more****, example, information, text,* ***should****, these, find, model, state, have, one, sentence, system, such, language, on, use,*
Corps	***observe****, other, task, base, if, two, than,* ***should****, datum, only, these, at, feature, also, set, all, have,* ***indicate****, one,* ***more****, note, each, predict, or, sentence,* ***may****, they, find, from, which,*
Conclusions	*note, than, information, only, each, but,* ***indicate****, et, predict, at, language, if, feature, such, these, al,* ***could****, also, other, sentence,* ***should****, one, state, have,* ***more****, they, or,* ***may****, from, on,*

TABLEAU 3.3 – Top 30 des mots-clefs les plus représentatifs du degré 1 selon les parties

Résumés	*chinese, use,* ***significant, compare, better, accuracy****, than, translation, be, by, two, benchmark,* ***can****, datum, for, english, both, base, evaluation, system, propose* ***improvement, significantly****, with, over, to, in,* ***improve, demonstrate****, baseline*
Introductions	*it, two, training, figure, dataset, accuracy, also, these, propose, which, work feature,* ***achieve****, section, word, system, how, language, find, datum, base, from experiment,* ***improve, demonstrate****, approach, task, paper, by, performance,*
Corps	*sentence, state, section,* ***best****, art, fig, task, feature, both, word,* ***outperform, can****, system, different, datum, experimental, accuracy, two, each, test, from, all, use, by, this, experiment, score, set, baseline, method,*
Conclusions	*word, than,* ***better****, also, different,* ***improvement****, have, over, from, both, experimental, by, set, this,* ***demonstrate****, base, datum, accuracy, system, approach,* ***outperform, report****, baseline,* ***achieve, improve****, dataset,* ***can****, task, use, method,*

TABLEAU 3.4 – Top 30 des mots-clefs les plus représentatifs du degré 2 selon les parties

FIGURE 3.2 – Diagramme en bâtons des spécificités des *claims* de degré 1

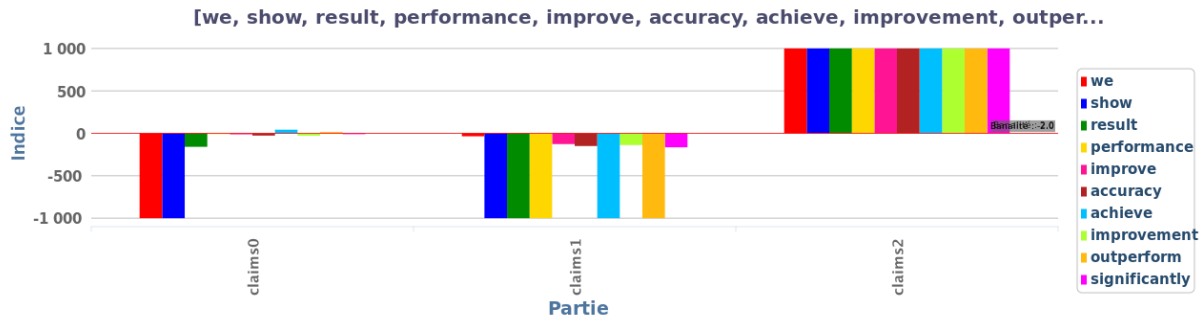

FIGURE 3.3 – Diagramme en bâtons des spécificités des *claims* de degré 2

Nous utilisons TXM pour visualiser les mots les plus spécifiques à chaque partition du corpus (une partition comprend tous les *claims* assignés à tel degré) et pouvons ainsi comparer ces spécificités avec les mots-clefs donnés par le *clustering* (voir Figures 3.2 et 3.3). Nous générons les diagrammes de spécificités de tokens avec des indices 1000 pour les catégories 1 et 2. Nous ne faisons pas celui de la catégorie 0, car, par définition, cette catégorie incertaine ne comprend pas de réel indice de modalité épistémique.

Pour la catégorie 1, moyennement certaine, nous retrouvons bien des indices qui semblent indiquer ce niveau de modalité épistémique : *can, it, find, may, more, indicate, could, appear, might, seem*. On s'intéresse aux mots qui ont des indices très faibles dans les autres partitions. C'est notamment le cas de *can, it, may*, qui ont des indices de -1 000 dans les degrés 0 et 2. En effet, ces deux modaux expriment une certitude assez modérée tandis que le pronom neutre permet une mise à distance, d'autant plus qu'il est beaucoup utilisé dans des phrases passives.

Quant à la catégorie 2, certaine, on retrouve encore des indices qui correspondent bien à l'étiquette : *we, show, result, performance, improve, accuracy, achieve, improvement, outperform, significantly*. Seul le mot *show* a un indice de -1 000 dans les deux autres catégories. Contrairement à ce que l'on pouvait penser, un seul autre token a un indice aussi faible pour la catégorie 0, *we*, pronom personnel largement utilisé pour réaliser des *claims*. Il semble alors que les catégories 1 et 2 soient plus éloignées, avec 4 tokens supplémentaires à -1 000, très représentatifs d'un haut degré de certitude vis à vis de résultats obtenus : *result, performance, achieve, outperform*.

Ces expériences nous permettent de voir que notre *clustering* et les étiquettes que nous avons attribuées aux différents *clusters* ainsi créés semblent pertinentes et justifiées.

Finalement, en écho au dernier paragraphe de la section 3.1, nous examinons les degrés de certitude qui ont été attribués aux *claims* précédemment donnés comme exemples et en ajoutons d'autres afin de représenter toutes les catégories par deux phrases :

- *« recent work on structure - aware model have show promising result on language modeling. »* [Zhang and Song, 2019] -> *degré 2*
- *« the experiment show that rcnn be very effective to improve the state of the art dependency parse on both english and chinese dataset »* [Zhu et al., 2015] -> *degré 2*
- *« in this paper we present a preliminary empirical study on whether and how much automatic grammatical error correction can help improve seq text generation »* [Ge et al., 2019] -> *degré 1*
- *« in order to be useful, a parser must be able to accept a wide range of input type, and must be able to gracefully deal with dysfluency, false start, and other ungrammatical input. »* -> *degré 1* [Malouf, 2000]
- *« in the past few year recurrent neural network rnn base architecture chopra et al gu et al nallapati et al see et al zhou et al li et al b a zhu et al have obtain state of the art result for text summarization. »* [Xu et al., 2020] -> *degré 0*
- *« these good result suggest that the learn representation capture linguistic and semantic property of the input that be relevant to the downstream re task a intuition that wa previously discuss for a variety of other nlp task by conneau et al »* [Alt et al., 2020] -> *degré 0*

D'après ces exemples et leur classification, il semblerait que notre *clustering* soit plutôt efficace et pertinent. En effet, les *claims* qui se voient assigner le degré 2 semblent très directs et assurés. C'est le verbe *show* qui est employé, on établit un rapport direct entre les données et les conclusions tirées. Bien que l'on emploie le même verbe dans le troisième exemple, la remarque est plus limitée et on parle d'amélioration d'un état de l'art particulier et non de tout un modèle voire un domaine. Quant aux exemples représentant le degré 0, on remarque que l'un est un simple commentaire neutre et factuel sur un article de leur état de l'art, tandis que l'autre prend la forme d'une suggestion, le degré de certitude exprimé est donc assez faible.

3.3.2 Validation de la forme des *clusters*

Afin d'avoir une meilleure idée des *clusterings* réalisés, nous utilisons des outils de visualisation qui montrent la répartition des *clusters* et ce, pour chaque partie, donc chaque *clustering* réalisé (voir Figures 3.4, 3.5, 3.6, 3.7) .

Attention, les légendes ne sont pas toujours identiques, les couleurs attribuées à chaque numéro de cluster *changent d'une figure à l'autre car nous avons ré-attribué manuellement les numéros afin que le degré 0 corresponde toujours à l'incertitude, le degré 1 à la certitude modérée et le degré 2 à la certitude*).

A partir de ces visualisations, on peut commencer à analyser et comparer les comportements des *claims*, ou plutôt, devrions-nous dire, des auteur·ices écrivant ces *claims*, selon les parties des articles.

Tout d'abord, on remarque que dans tous les cas, les *clusters* restent relativement homogènes et collés, ils ne forment qu'un seul gros bloc, les séparations ne sont pas nettes. Cela témoigne encore une fois de la complexité de la tâche, et justifie que certains résultats des évaluations de nos apprentissages non-supervisés soient assez bas.

On voit aussi que le *cluster* correspondant au degré 1, soit le degré moyen, se trouve toujours au milieu des deux autres, ce qui témoigne en effet d'une gradation entre les différentes catégories. Ainsi, les degrés 0 et 2 se retrouvent plus marqués et chacun d'un côté des graphiques, ne se touchant que sur la figure des corps.

Néanmoins, des tendances plus propres à chaque partie se dégagent, et cette remarque est plus ou moins vraie selon les cas. Par exemple, les *clusters* les mieux distingués et séparés sont ceux des conclusions, ce qui explique également que ce soit la partie qui obtienne les meilleurs scores d'évaluation. A l'inverse, on peut penser que si on avait obtenu les résultats pour les corps, ceux-ci auraient été les plus bas car les *clusters* sont très mélangés.

Nous pourrions supposer que les *claims* présents dans les corps d'articles sont très homogènes et adoptent toujours le même degré de certitude, ce qui rend leur classification plus délicate. Par contre, les tendances sont plus tranchées en conclusions, où l'on a d'un côté des *claims* faisant preuve d'une forte certitude, ou au contraire, d'une forte incertitude. Les *claims* des résumés et des introductions semblent incarner un cas de figure plus mesuré, entre ces deux tendances.

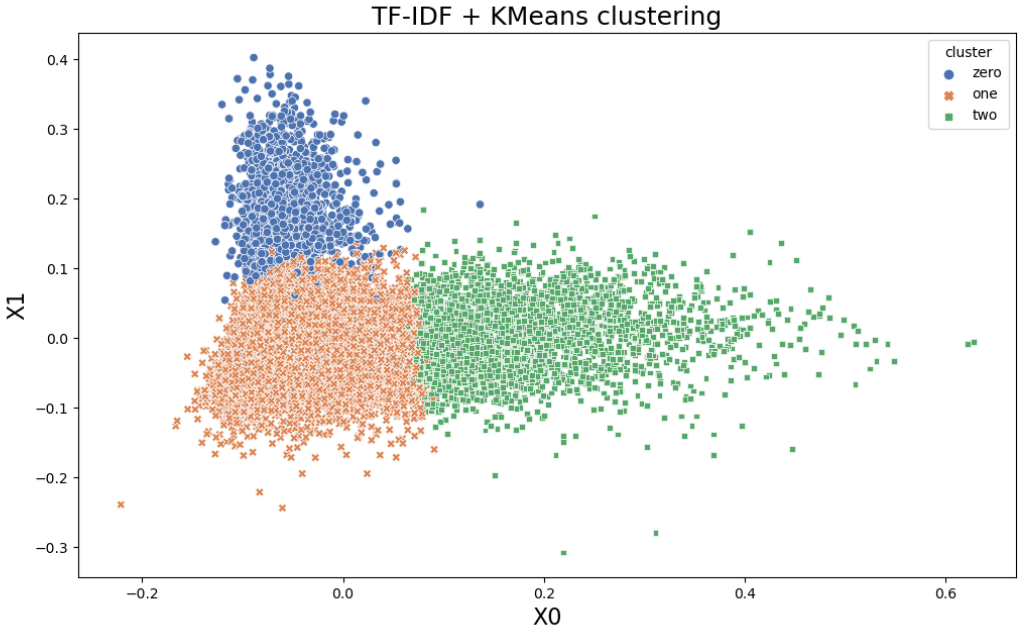

FIGURE 3.4 – Visualisation du *clustering* sur les résumés

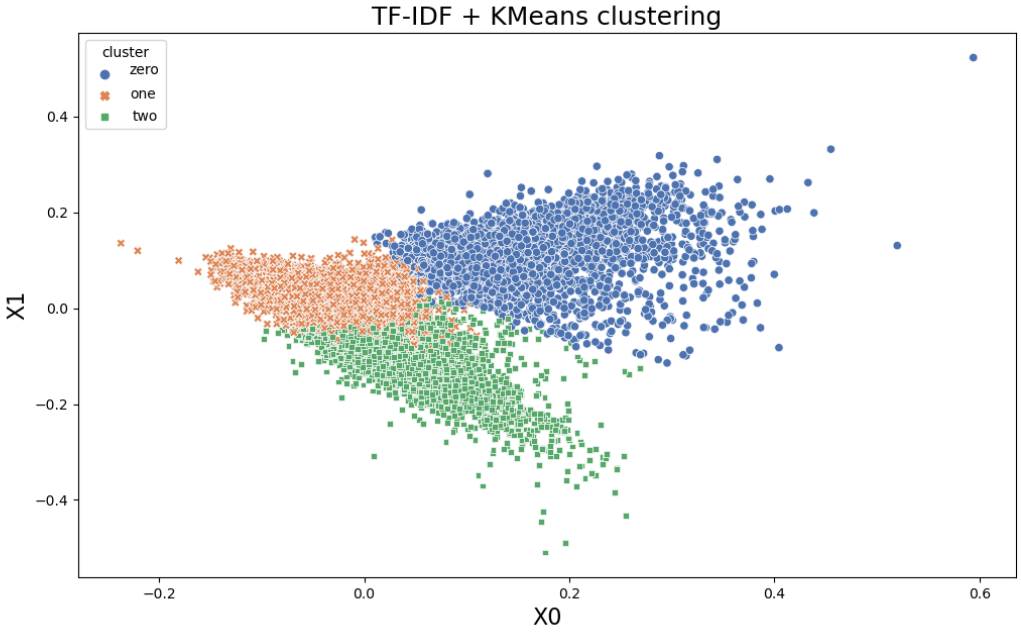

FIGURE 3.5 – Visualisation du *clustering* sur les introductions

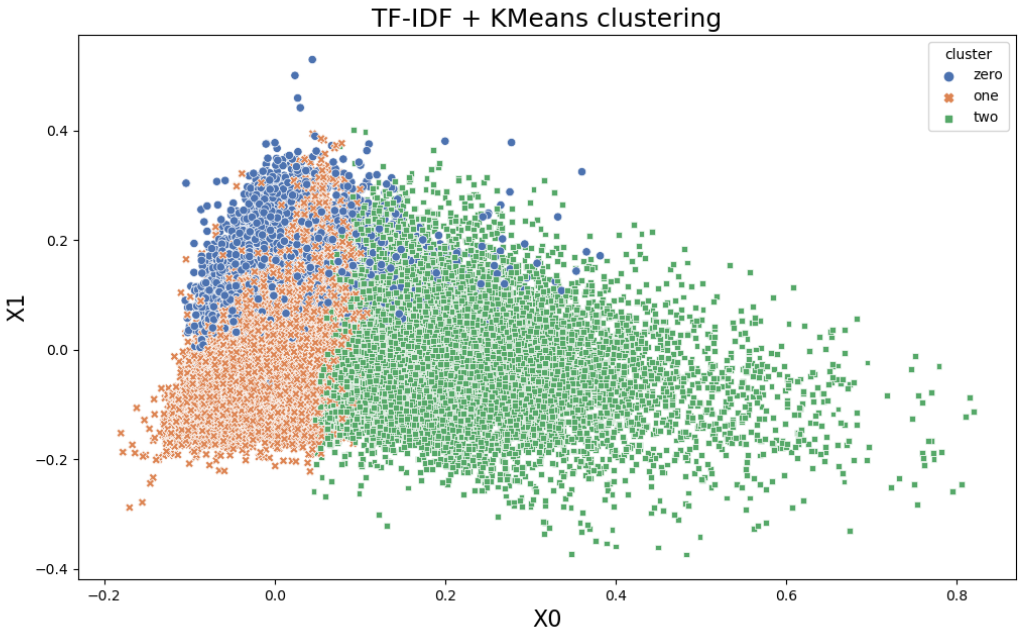

FIGURE 3.6 – Visualisation du *clustering* sur les corps d'articles

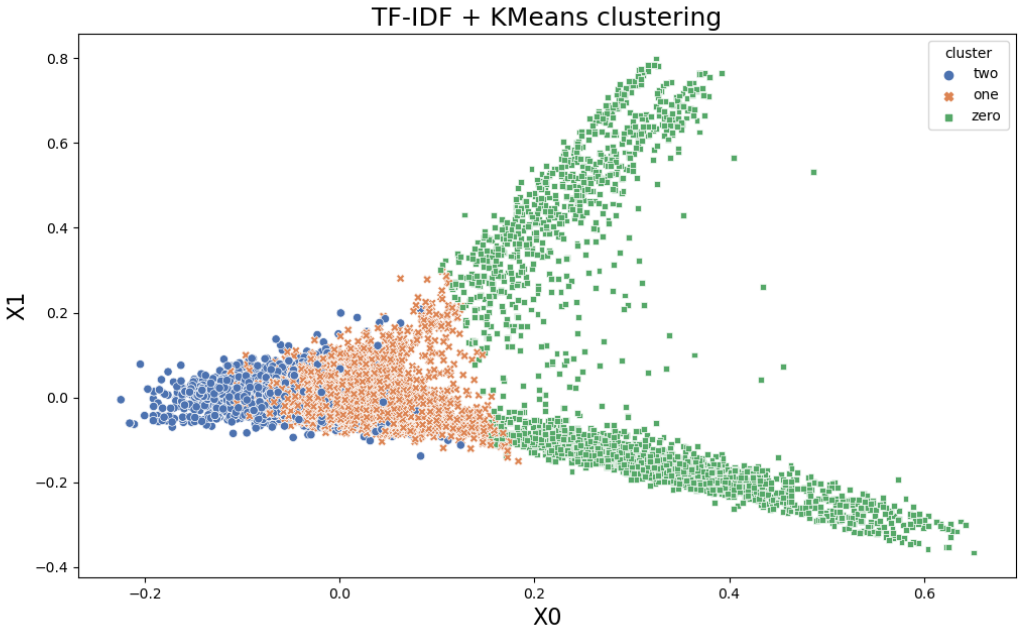

FIGURE 3.7 – Visualisation du *clustering* sur les conclusions (attention : l'axe des Y est différent des figures précédentes)

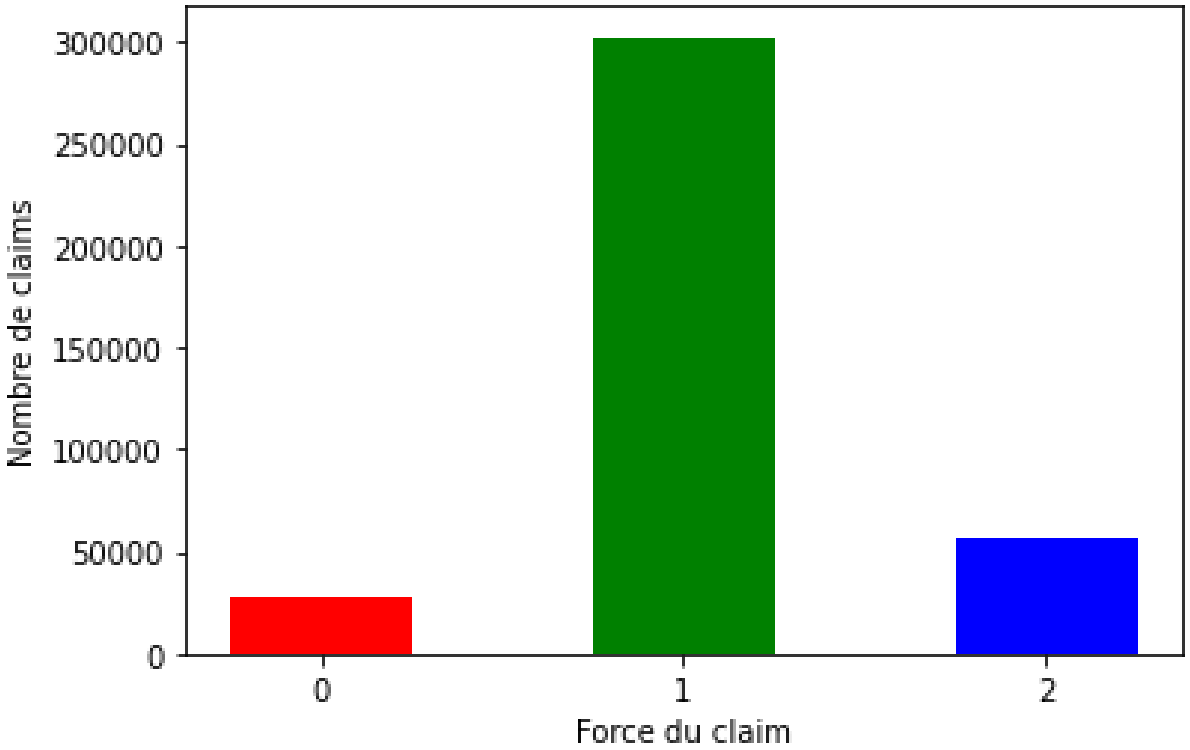

FIGURE 3.8 – Nombre absolu de *claims* par degré de force, sans prendre en compte les parties

3.4 Analyse des clusterings

3.4.1 Comparaison globale

Nous avons ensuite utilisé nos résultats de *clustering* pour mener de plus larges investigations. Nous commençons par calculer la proportion d'articles contenant au moins un *claim* dans le corpus : celle-ci s'élève à 98,9 % (6 305 articles contenant au moins un *claim* sur un total de 6 372 articles dans le corpus). Ensuite, nous comptons le nombre de *claims* appartenant à tel ou tel degré de force en concaténant les différents fichiers CSV contenant les résultats des différents *clusterings* (voir Figure 3.8). On voit alors que c'est le degré de force 1 qui est largement le plus représenté. Cela n'est pas nécessairement surprenant, car il correspond à la catégorie la plus nuancée, qui n'exprime ni l'incertitude ni la certitude totales. Dans une moindre mesure, ce sont ensuite les claims de degré 2, associés à une certitude élevée, qui sont plus nombreux que ceux associés au degré 0.

Nous nous intéressons ensuite aux parties contenant le plus de *claims* à travers la figure 3.9. Nous pouvons y voir que, comme on pouvait s'y attendre, on trouve le plus de *claims* dans les corps d'articles, puisque ce sont les parties les plus longues des articles, comptant donc le plus de phrases et donc potentiellement le plus de *claims*. Néanmoins, la figure 3.10, qui utilise les moyennes, montre les mêmes tendances, la longueur ne semble donc pas avoir une si grande influence.

Les *claims* sont ensuite les plus nombreux en conclusions. Nous supposons que cela est dû à la nature même de cette section, dont le but est de récapituler les résultats obtenus et les conséquences que les auteur·ices en tirent. Finalement, les introductions ont un

nombre absolu de *claims* plus élevé que les résumés. Cela pourrait s'expliquer du fait que les introductions contiennent souvent un bref état de l'art dans lequel les auteur·ices reprennent les *claims* les plus saillants des articles cités.

Ensuite, nous recoupons les informations afin de regarder de plus près le nombre de *claims* appartenant à tel degré dans telle partie (voir Figure **??**).

On observe un comportement très différent selon la force de l'affirmation. On retrouve la forte prévalence de ceux présents dans les corps d'articles. Cette prévalence est même écrasante dans le cas des affirmations moyennement certaines dans les corps d'articles. En effet, on peut penser que les affirmations qui se situent dans cette partie correspondent plutôt à des observations qui se veulent assez neutres et répondant aux expériences menées et présentées.

On retrouve ainsi la forte présence des *claims* de degré 1, qui est la catégorie d'affirmations majoritaire dans toutes les parties d'articles. On peut noter que les conclusions semblent utiliser presque uniquement des affirmations moyennement certaines, et quasiment aucune très certaines. Cela est plus nuancé et légèrement inversé pour les résumés et les introductions, qui présentent plus de *claims* de force 2 que de force 0.

3.4.2 Progression intra-article

Nous poursuivons notre étude en nous intéressant à la manière dont les degrés des affirmations évoluent dans les articles (voir Figure 3.12). Pour cela, nous prenons en compte les articles qui présentent au moins un *claim* dans le résumé ou l'introduction, au moins un dans le corps, et au moins un en conclusion.

Nous mettons ensuite dans la catégorie *progression* ceux dans lesquels les affirmations sont de plus en plus certaines. Cela concerne par exemple, une affirmation classée comme force 0 en introduction, puis force 1 dans le corps, et finalement force 2 dans la conclusion. La catégorie *régression* regroupe quant à elle les articles qui suivent une tendance inverse avec des affirmations aux degrés plus forts au début qu'à la fin. Nous créons également des catégories dédiées aux cas où l'on observe une régression puis une progression (type affirmation de force 2 en introduction, puis 1 dans le corps, puis 2 à nouveau dans la conclusion), et une catégorie progression puis régression pour le cas inverse. En parallèle, nous prenons en compte les cas des articles qui ne présentent aucun changement dans le degré d'affirmation utilisé et regroupons les articles où les affirmations sont toujours de force 0, ceux où elles sont toujours de force 1 et celles qui restent au degré 2.

De manière assez rassurante et cohérente, on constate que la majorité des articles suivent une progression, commençant donc par énoncer des affirmations incertaines puis devenant de plus en plus certaines au fur et à mesure de leur étude. Ainsi, ce sont 841 articles qui connaissent une augmentation de leur degré de certitude au fil de l'article. Cependant, la catégorie *régression puis progression* est également bien représentée : 334 articles suivent cette tendance.

Les cas de *régression* ou de *progression puis régression* sont peu nombreux : 137 (55 + 82) articles correspondent à l'une ou l'autre de ces catégories. Cela va de paire avec nos observations selon lesquelles les conclusions ont très peu de *claims* de force 2.

La catégorie la plus représentée pour les cas où les *claims* restent stables est alors celle des articles n'ayant que des *claims* de force 1 ; 536 articles sont dans ce cas. De même que précédemment, on peut attribuer cela au fait qu'il s'agit de la catégorie la plus nuancée.

Il y a également peu d'articles n'ayant que des affirmations de force 0 ou de force 2 : ces cas de figure représentent 137 articles (8 + 129). En effet, n'avoir que des affirmations

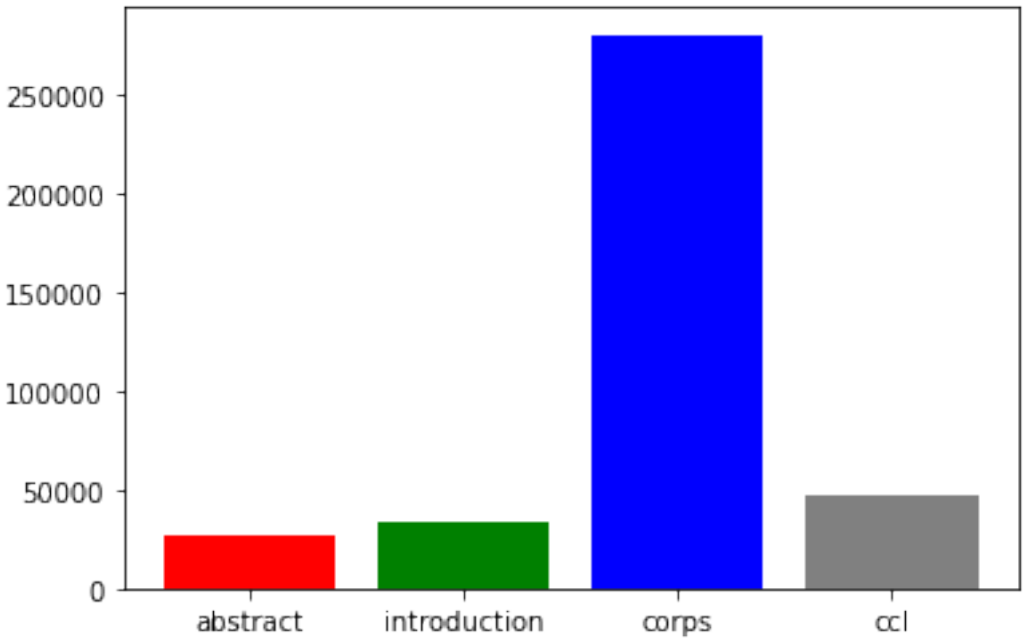

FIGURE 3.9 – Nombre absolu de *claims* contenus par partie

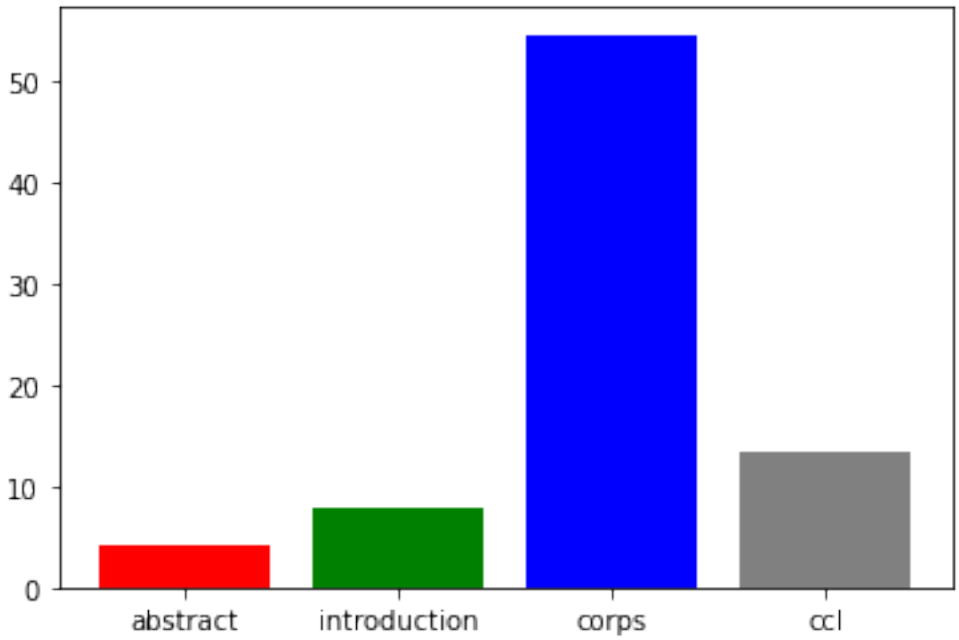

FIGURE 3.10 – Nombre moyen de *claims* par article selon la partie

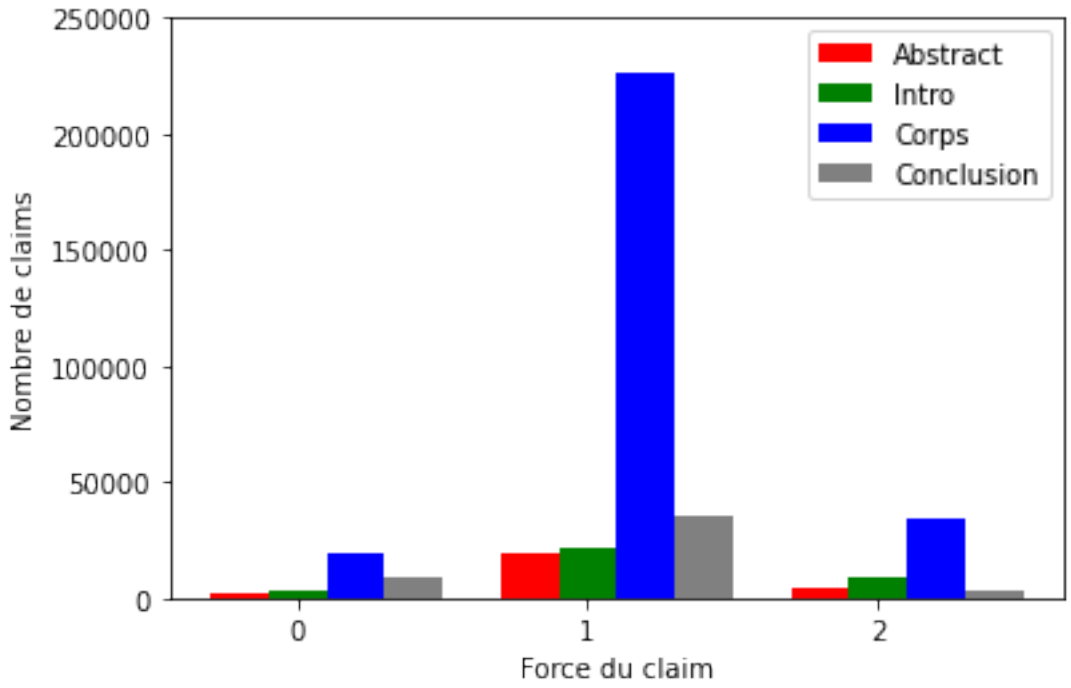

FIGURE 3.11 – Nombre de *claims* par catégorie selon la partie

incertaines reviendraient à admettre que son étude n'apporte rien tandis que n'en avoir que des très certaines pourrait s'avérer pompeux (or comme nous l'avons vu dans notre état de l'art, l'humilité est une valeur cruciale en science).

3.4.3 Évolution diachronique

Finalement, nous nous intéressons à l'évolution du nombre de *claims* au fil du temps, mais aussi à leur localisation dans l'article et à leur force afin de remarquer d'éventuelles tendances selon les années.

Sur les figures 3.13 et 3.14, nous nous intéressons uniquement aux nombres de *claims* présents dans les articles et les regroupons par année.

Il nous permet de visualiser une certaine stabilité entre 1979 et 2005, suivie d'une progression ascendante croissante d'année en année, qui explose en 2019 et 2020. Ainsi, la figure sur le nombre absolu prouve la forte augmentation du nombre de *claims* au fil des années, qui s'explique notamment par la forte augmentation du nombre d'articles publiés en général (le nombre de publications par année est compris entre 25 et 778, voir Figure 3.15).

Afin de mener une comparaison équilibrée, nous calculons le nombre moyen de *claims* contenus par article selon les années (voir Figure 3.14). Nous observons une augmentation du nombre moyen de *claims* par article entre 1979 et 1989, suivie d'une période de baisse de ces nombres moyens entre 1990 et 2000. Cette baisse pourrait être expliquée par la diminution du nombre de publications pour ces années, mais également par le changement de paradigme qui survient à cette période. Il s'agit en effet du début du *machine learning*, on peut émettre l'hypothèse d'un manque de sûreté vis à vis de ces méthodes. La tendance repart ensuite à la hausse jusqu'en 2020, où elle atteint son maximum. On peut également observer des nombres quasiment inédits tous les ans à partir de 2010.

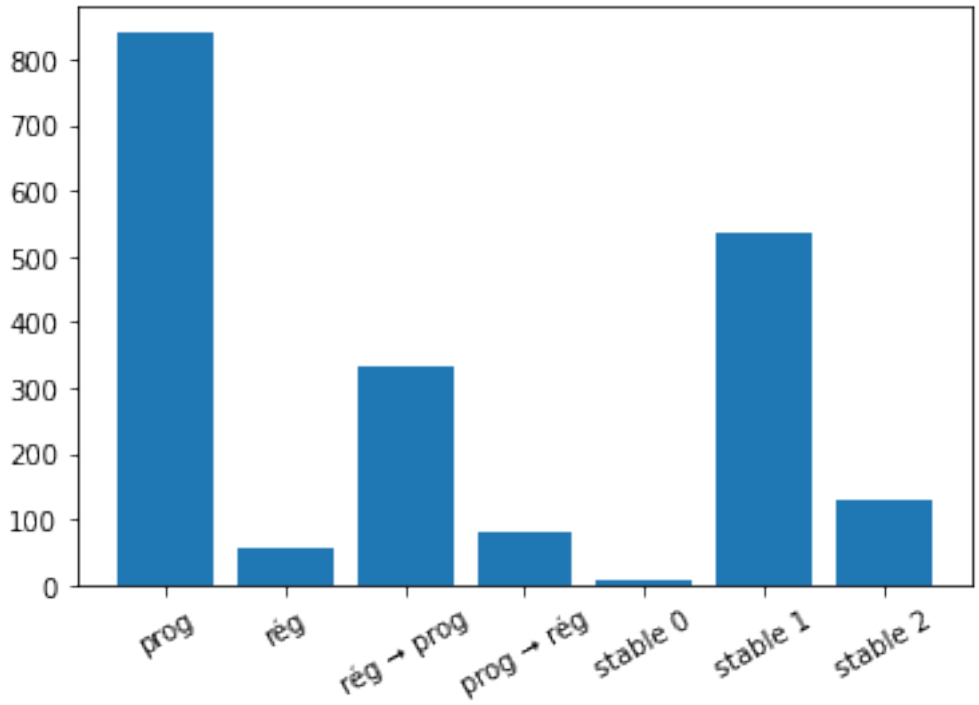

FIGURE 3.12 – Nombre d'articles suivant telle évolution dans la force des affirmations au gré des parties (prog : progression, rég : régression, -> : puis)

Les années 2010 constituent en effet un tournant pour notre domaine. C'est au début de cette décennie que l'on voit l'apparition et la popularisation d'outils de TAL comme Siri (par Apple en 2011), Cortana (par Microsoft en 2014), Alexa (par Amazon en 2014) mais également du *deep learning* et des *embeddings* (Word2Vec présenté pour la première fois en 2013 [Mikolov et al., 2013]) [7]. Le TAL se démocratise et suscite l'intérêt de la communauté scientifique, mais également des investisseurs.

Ensuite, nous observons l'évolution de la répartition des *claims* en fonction de leur degré (voir Figure 3.16) et de leur localisation dans l'article (voir Figure 3.17).

Nous remarquons que la proportion de *claims* moyennement certains (de degré 1) diminue au fil du temps, bien qu'elle reste beaucoup plus élevée que les autres. À l'inverse, les proportions de *claims* incertains (degré 0) et certains (degré 2) augmentent progressivement. Ces deux courbes restent proches l'une de l'autre, bien que l'on observe une augmentation légèrement plus élevée pour les *claims* certains.

Le graphique 3.17 indique que la proportion de *claims* présents dans le corps a eu tendance à augmenter avec le temps tandis que ceux présents dans les autres parties ont diminué. Nous pouvons penser que la croissance de l'importance du corps d'article est due à l'augmentation de la longueur de cette partie. En effet, entre 1979 et 1999, les corps d'articles contiennent en moyenne 122 phrases tandis que cette longueur moyenne atteint 227 pour les corps des articles publiés entre 2000 et 2020. Cela pourrait s'expliquer par l'augmentation du nombre d'expériences dans les articles.

De même, le nombre de *claims* a augmenté dans les introductions et les conclusions. Cependant, cette augmentation est légère, surtout pour les conclusions. La longueur moyenne

7. https://en.wikipedia.org/wiki/Natural_language_processing#Neural_NLP_(present)

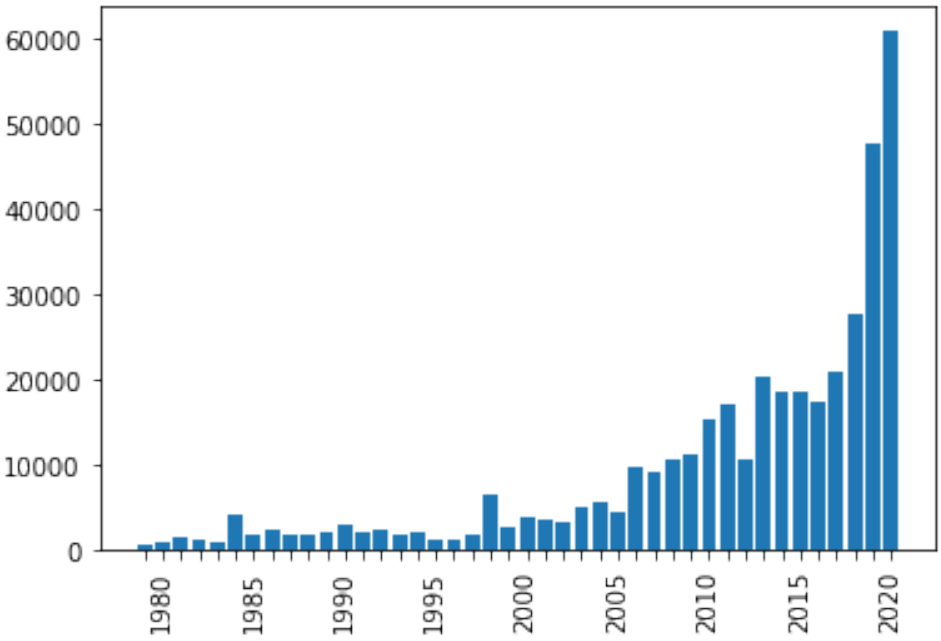

FIGURE 3.13 – Évolution du nombre absolu de *claims* selon les années

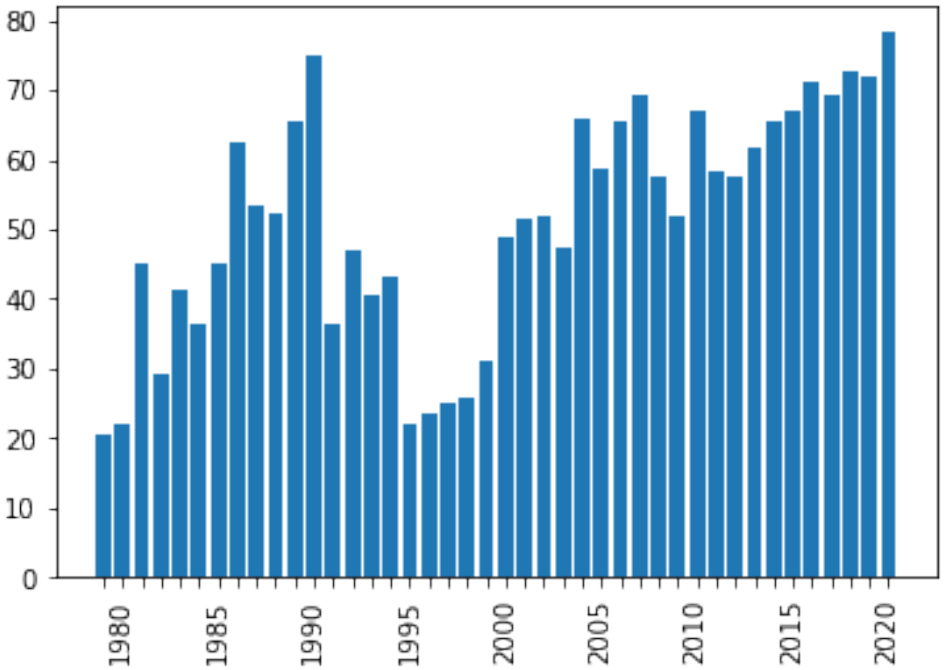

FIGURE 3.14 – Évolution du nombre moyen de *claims* par article selon les années

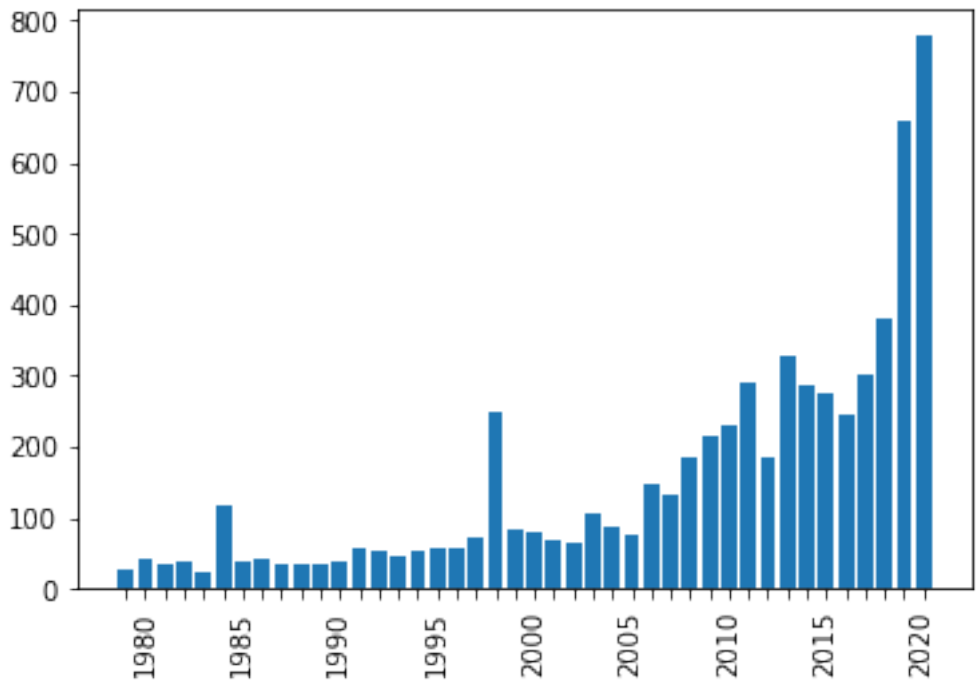

FIGURE 3.15 – Évolution diachronique du nombre de publications à ACL par année

de phrases dans ces parties a diminué, cela peut donc constituer une explication. Nous pouvons penser que cette diminution du nombre de *claims* et du nombre de phrases dans ces parties est une conséquence directe de l'augmentation de la longueur des corps d'articles : pour compenser, il faut raccourcir les autres parties.

Le cas des résumés (*abstract*) est légèrement différent : la présence des *claims* dans cette partie a d'abord augmenté avant de grandement diminuer à partir de 2000. Cela peut s'expliquer par une raison similaire à celle évoquée pour les corps d'articles : à l'inverse, la longueur moyenne des résumés a augmenté puis diminué au fil du temps, passant d'une moyenne de 21 phrases avant 2000 à une moyenne de 14 après 2000. La diminution de la longueur des résumés pourrait être due à des contraintes imposées par la conférence.

Nous pourrions également penser que la forte augmentation du nombre d'articles publiés à ACL pousse les auteur·ices à attirer plus rapidement l'attention du lectorat, qui a une masse de résumés à disposition et n'a pas nécessairement le temps ou l'envie d'en lire des centaines, d'autant plus s'ils sont longs. Les auteur·ices ont donc tout intérêt à être plus synthétiques dans leur résumé et à le raccourcir au maximum.

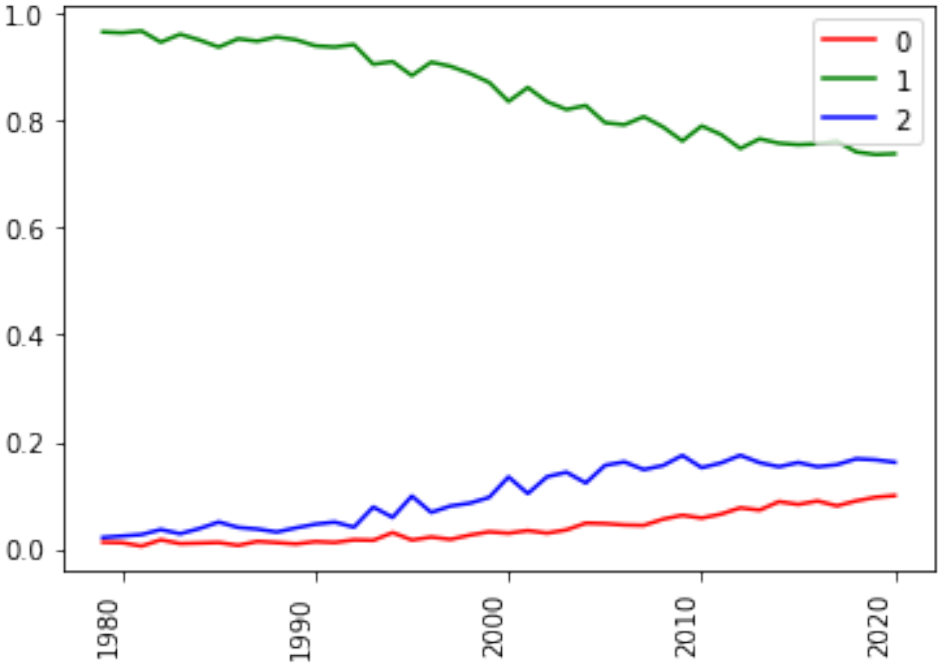

FIGURE 3.16 – Évolution diachronique de la répartition des *claims* selon le degré

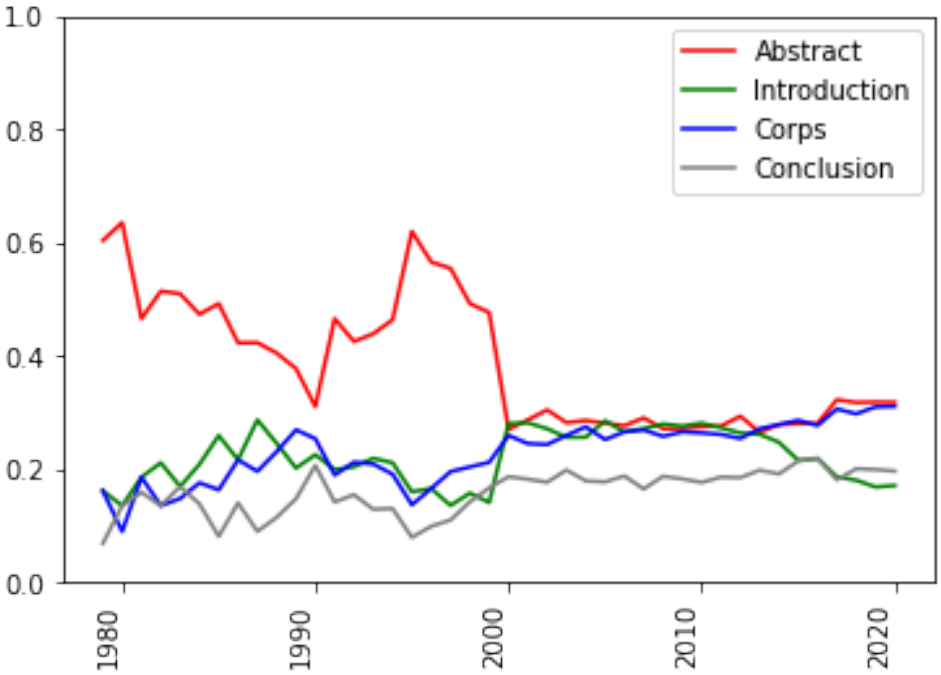

FIGURE 3.17 – Évolution diachronique de la répartition des *claims* selon la partie

Chapitre 4

Corrélation avec certains facteurs sociologiques

Les résultats généraux que nous obtenons sur notre corpus semblent assez homogènes, aucune catégorie ne semble en émerger. Afin d'y remédier, nous testons plusieurs hypothèses en utilisant des caractéristiques propres aux auteur·ices pour voir si celles-ci sont corrélées avec le nombre et la qualité des *claims*. Nous choisissons alors de nous intéresser à la potentielle corrélation avec le genre de l'auteur·ice, de son continent de rattachement et de ses éventuelles affiliations à des institutions de prestige. Ces critères ont été sélectionnés car les informations qui permettent de les détecter sont présents directement dans les en-têtes des articles et recoupent des catégories déjà existantes ainsi que d'autres hypothèses présentées dans d'autres travaux (par exemple, l'hypothèse selon laquelle les femmes seraient moins assertives que les hommes que nous avons détaillée dans la section 2.4.1). Nous présentons les résultats obtenus suite à ces expériences dans ce chapitre.

4.1 Corrélation avec le genre

Comme précisé dans la partie méthodologie dédiée (Section 2.4.1), seuls 69,5 % des articles sont traités ici car ce sont ceux auxquels on a pu associer leur genre aux auteur·ices.

Rappelons également que parmi 14 198 prénoms détectés dans les articles et ayant pu être rattachés à un genre, plus de 79 % sont masculins (11 132/14 198) et donc un peu plus de 21 % sont féminins (3 066/14 198). Cela témoigne d'un déséquilibre encore bien présent dans la communauté scientifique avec une sous-représentation des femmes dans la recherche.

Ces données coïncident avec celles obtenues par Mohammad [2020]. Son étude porte sur 44 894 articles de conférences de TAL (dont ACL) publiés entre 1965 et 2019. Il y indique qu'environ 30 % des auteur·ices de son corpus sont des femmes, ce qui correspond exactement à la proportion de femmes dans la science qu'il présente. Ce chiffre descend néanmoins à 24,5 % s'il prend seulement en compte les articles d'ACL et s'approche du résultat que nous avons obtenu. D'après ses expériences, le nombre de papiers ayant une première autrice s'élève à 29 %, et ceux ayant une dernière autrice à 25 %. Mohammad [2020] développe également les enjeux et les conséquences de cet écart entre les genres : en outre d'être injuste, ce déséquilibre amoindrit la productivité, le bien-être, la prise de décisions, et même la stabilité politique et économique. Le but de son article est alors de susciter une prise de conscience afin d'améliorer l'état actuel de la situation.

Évolution du nombre de femmes à ACL

Nous nous arrêtons un instant sur cette question et réalisons des graphiques montrant l'évolution dans le temps du nombre d'auteurs et d'autrices publié·es à ACL. Nous prenons toutes les personnes mentionnées dans les listes auteur·ices des articles (voir Figure 4.1) ;

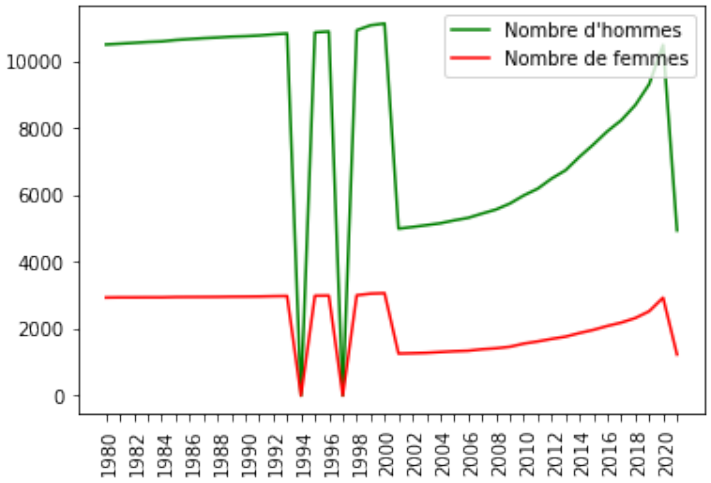

FIGURE 4.1 – Evolution diachronique du nombre d'auteurs et d'autrices publiés à ACL (attention, l'échelle n'est pas la même que sur les graphiques suivants car nous nous intéressons ici au nombre de personnes et non pas d'articles)

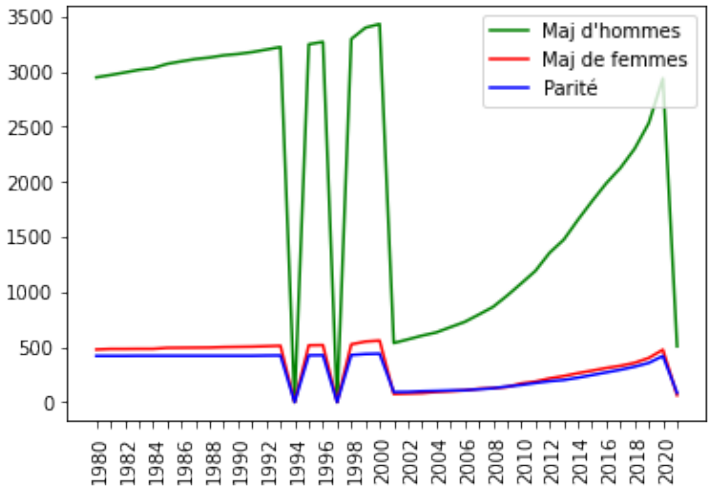

FIGURE 4.2 – Evolution diachronique du nombre d'articles écrits par des majorités d'hommes, de femmes ou des parités

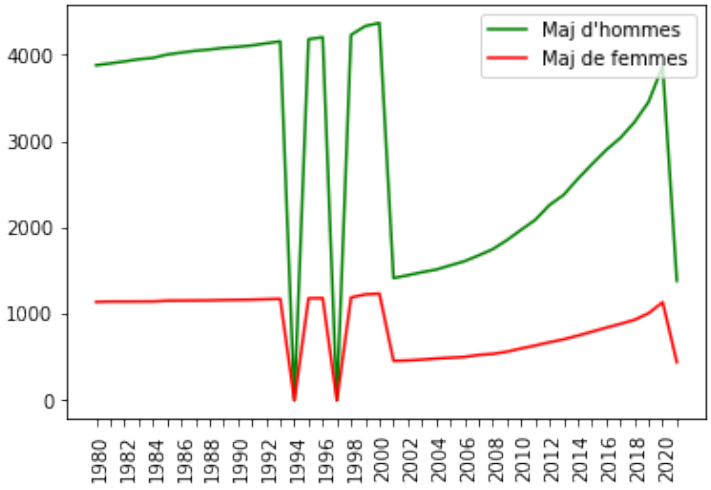

FIGURE 4.3 – Evolution diachronique du nombre d'articles ayant des premiers auteurs ou des premières autrices

le but ici n'est pas d'avoir les nombres uniques de noms mais bien les nombres d'articles signés par des hommes ou des femmes, notre liste contient donc évidemment des doublons, mais ce n'est pas ici ce qui nous intéresse. Nous générons également un graphique représentant le nombre d'articles écrits en majorité par des hommes, des femmes ou une parité (voir Figure 4.2), et enfin du nombre d'articles comprenant un premier auteur face à ceux ayant une première autrice (voir Figure 4.3). Ainsi, l'écart entre le nombre d'hommes et de femmes est encore plus flagrant, les courbes concernant les femmes sont toujours beaucoup plus basses que celles des hommes, et, contrairement à ce que l'on aurait pu croire, la proportion de femmes, qui était pourtant au départ assez stable, semble baisser depuis 2000.

Afin d'encore mieux rendre compte de cela, on s'intéresse seulement à la proportion d'autrices publiées à ACL et on calcule des proportions. Cela donne la figure 4.4. Toutefois, cette figure est biaisée car les valeurs qui sont à zéro en 1993 et 1997 correspondent en fait à des valeurs manquantes pour ces années, et non pas à des années où il n'y aurait aucune autrice. On utilise alors une méthode de remplissage afin de combler les trous causés par les valeurs que nous n'avons pas pu obtenir, ce qui permet également de voir de plus près l'évolution (voir Figure 4.5). Il faut garder à l'esprit que ce graphique est très zoomé, l'évolution est en fait très peu importante, on reste toujours entre 20 et 22 % d'autrices selon les années, ce qui témoigne d'une assez grande stabilité. Mais, contrairement à ce que l'on aurait pu penser (ou espérer), on perçoit une très légère tendance à la décroissance.

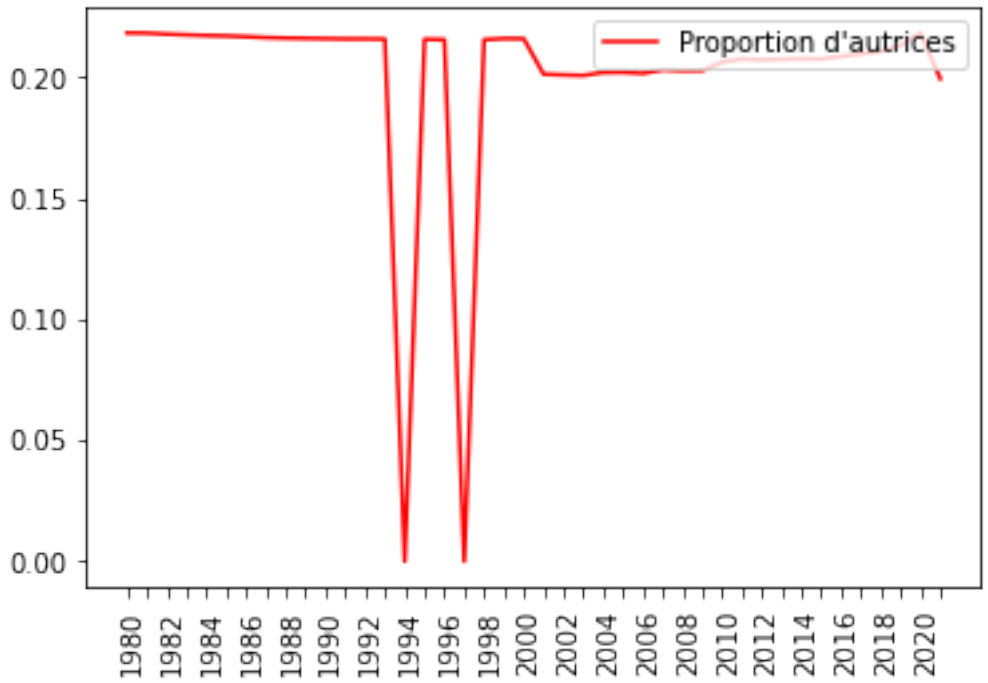

FIGURE 4.4 – Evolution diachronique du nombre d'autrices publiées à ACL, en proportions

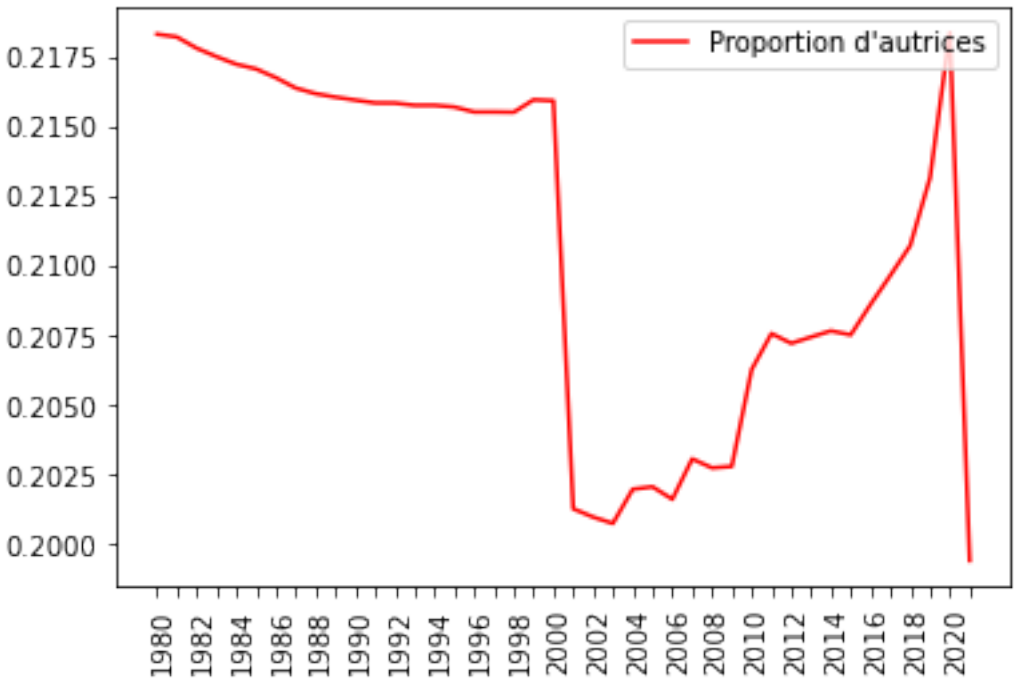

FIGURE 4.5 – Evolution diachronique du nombre d'autrices publiées à ACL, en proportions avec remplissage

Comparaison des *claims* selon le genre

Nous revenons ensuite à notre sujet initial et faisons deux pré-traitements différents, menant chacun à des analyses et comparaisons spécifiques.

Tout d'abord, nous décidons de dire qu'un article est catégorisé comme écrit par des hommes ou par des femmes ou par une égalité d'hommes et de femmes en prenant en compte le genre majoritaire du groupe d'auteur·ices. Dans ce cas, nous décomptons 3 433 (soit 77 %) d'articles écrits majoritairement par des hommes, 559 (soit 12 %) par des femmes et 440 (soit 10 %) par autant d'hommes que de femmes.

Nous comptons ensuite le nombre de *claims* présents dans chaque article associé à tel ou tel genre et en faisons des graphiques à partir des chiffres absolus d'une part (voir Figure 4.6), et relatifs de l'autre (en divisant le nombre de claims par le nombre d'articles associés au genre en question, ce qui revient à donner le nombre de *claims* ou plus précisément d'indices de *claims* associés par article associé à ce genre, voir Figure 4.7). Les chiffres absolus ne font que refléter les proportions d'articles appartenant à tel ou tel genre, c'est pourquoi l'utilisation de nombres relatifs est plus pertinente. On y observe que ce sont les groupes d'auteur·ices parfaitement mixtes qui émettent le plus de *claims* dans leurs articles, suivis par les groupes constitués majoritairement d'hommes puis de femmes.

On s'intéresse ensuite à la répartition des *claims* selon leur force pour chacun de ces genres (voir Figure 4.9). Grâce aux nombres relatifs et à l'inverse de la représentation avec nombres absolus qui pourrait nous induire en erreur, on se rend compte que les tendances sont extrêmement similaires (mais les chiffres exacts sont bien différents, il ne s'agit donc pas d'une erreur), le genre ne semble pas influer, c'est toujours la force 1 qui est la plus représentée.

De la même manière, on prend cette fois-ci en compte la partie où le *claim* est effectué (voir Figures 4.10 et 4.11). Là encore, on remarque que les proportions sont très similaires. Peu importe le genre majoritaire, les *claims* sont toujours les plus présents dans les corps d'articles, puis dans une bien moindre mesure en conclusion, puis en introduction et dans les résumés.

Nous réitérons toutes ces expériences avec notre deuxième approche qui, cette fois-ci, considère qu'un article est écrit par une femme si la première personne du groupe d'auteur·ices a été identifiée comme telle, et inversement pour les hommes ; la catégorie égalité n'est donc plus utilisée.

Nous comptons alors 3 488 dont les premiers auteurs sont détectés comme hommes (soit plus de 78 %) et 944 (soit environ 21 %) comme femmes. Proportionnellement à cela, nous voyons à quel point l'écart parait immense sur la figure 4.12, mais légèrement nuancé quand on s'intéresse aux nombres relatifs. Il reste quand même significatif : en moyenne, on trouve plus de 60 *claims* (ou plus précisément, indices de *claims*) par article dont l'auteur principal est un homme contre seulement une quarantaine par article avec une première autrice. Cela semble confirmer les études citées en début de section, les femmes émettraient donc bien moins de *claims*, ce qui pourrait avoir des causes sociologiques et psychologiques. Ces causes peuvent par exemple être liées à une confiance en soi moins élevée et un potentiel syndrome de l'imposteur, dans ce cas syndrome de l'impostrice ([Young, 2011] et [Paterson and Vincent-Akpu, 2022]), qui s'explique notamment par la faible proportion de femmes dans le domaine, comme nous l'avons vu plus haut.

Sur les figures 4.14 et 4.15, on voit que, malgré une différence quantitative importante,

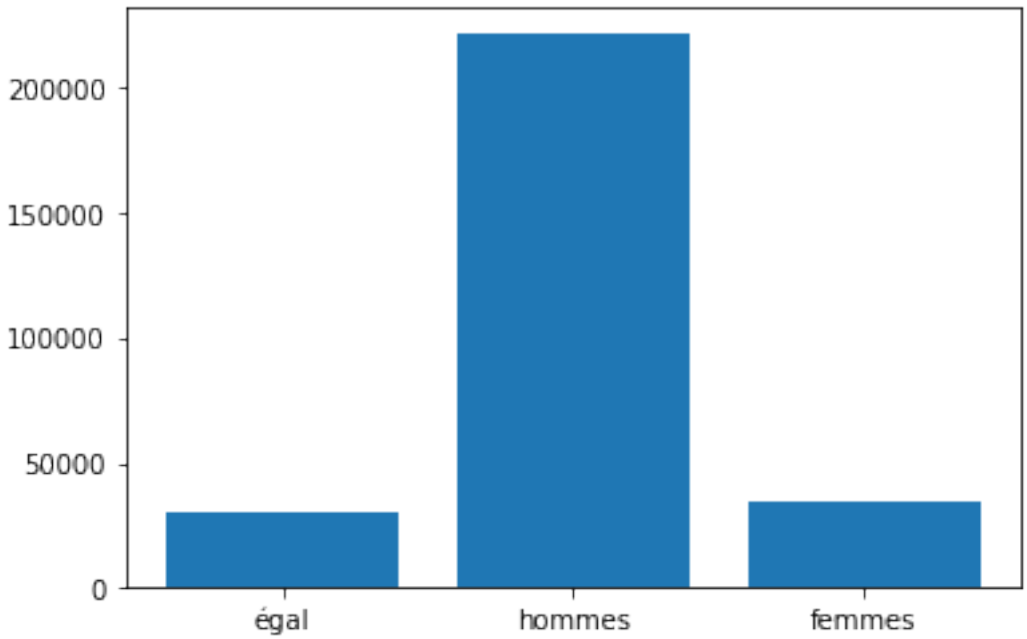

FIGURE 4.6 – Nombres absolus de *claims* selon le genre majoritaire du groupe d'auteur·ices de l'article

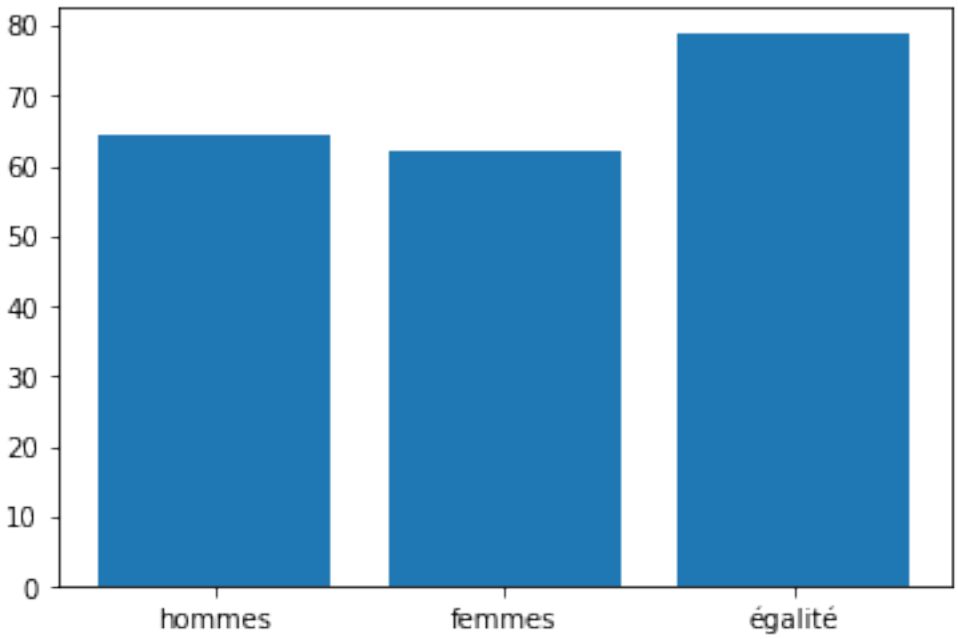

FIGURE 4.7 – Nombre moyen de *claims* par article selon le genre majoritaire du groupe d'auteur·ices de l'article

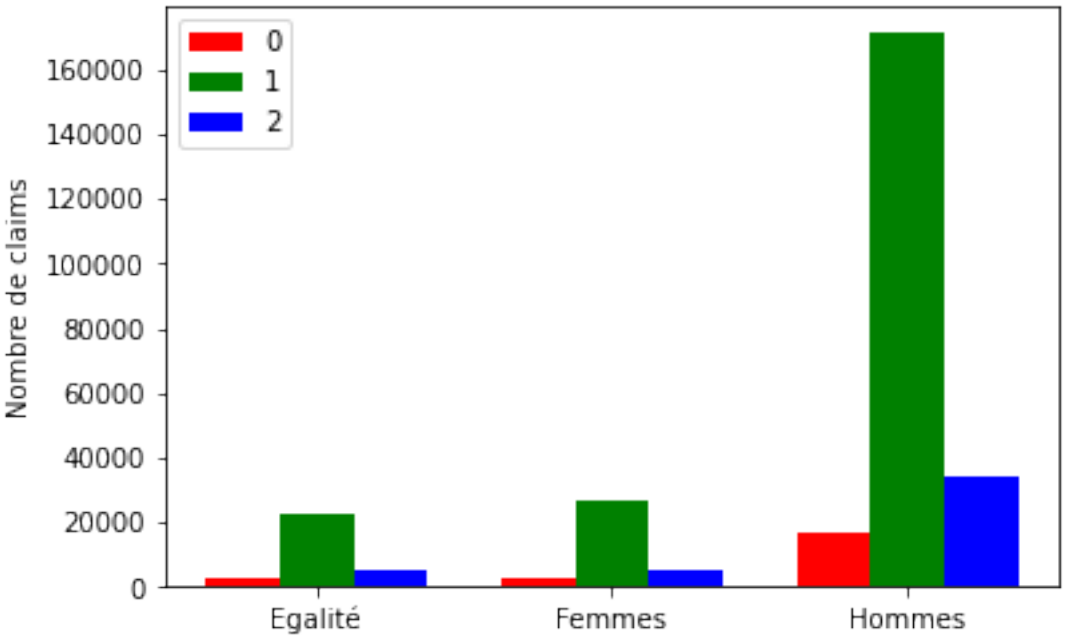

FIGURE 4.8 – Nombre absolu de *claims* par genre majoritaire selon la force

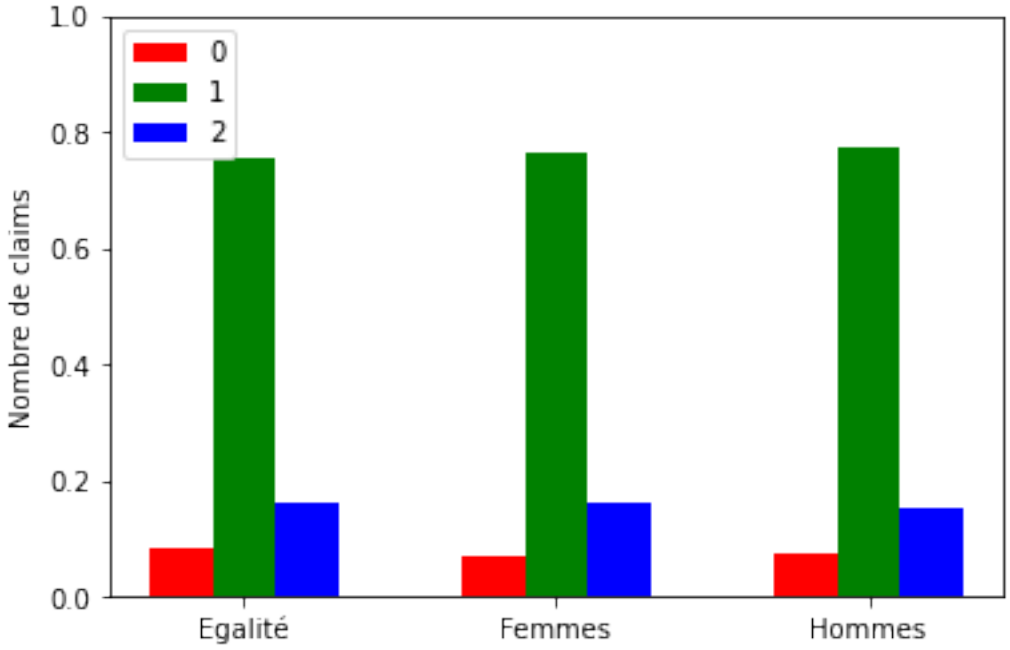

FIGURE 4.9 – Proportion de *claims* par genre majoritaire selon la force

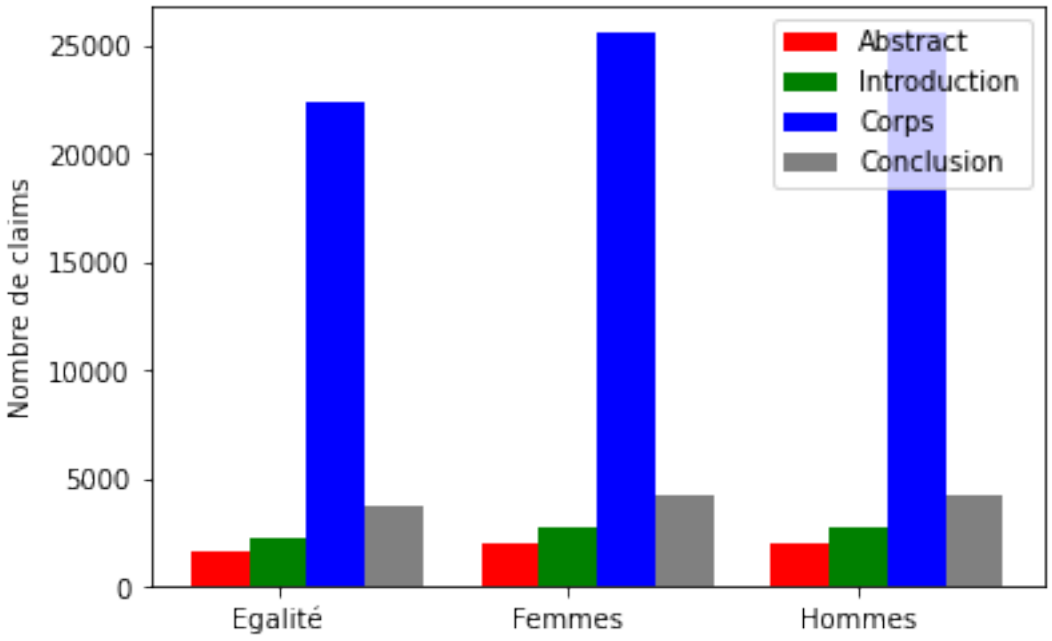

FIGURE 4.10 – Nombre de *claims* par force selon les parties et le genre majoritaire

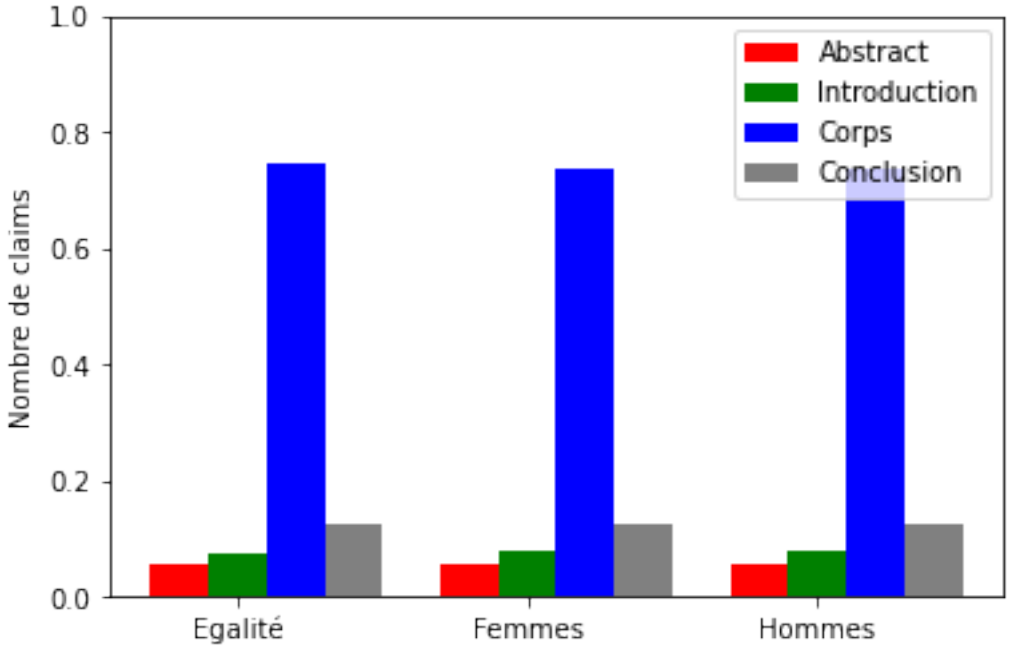

FIGURE 4.11 – Proportion de *claims* par force selon les parties et le genre majoritaire

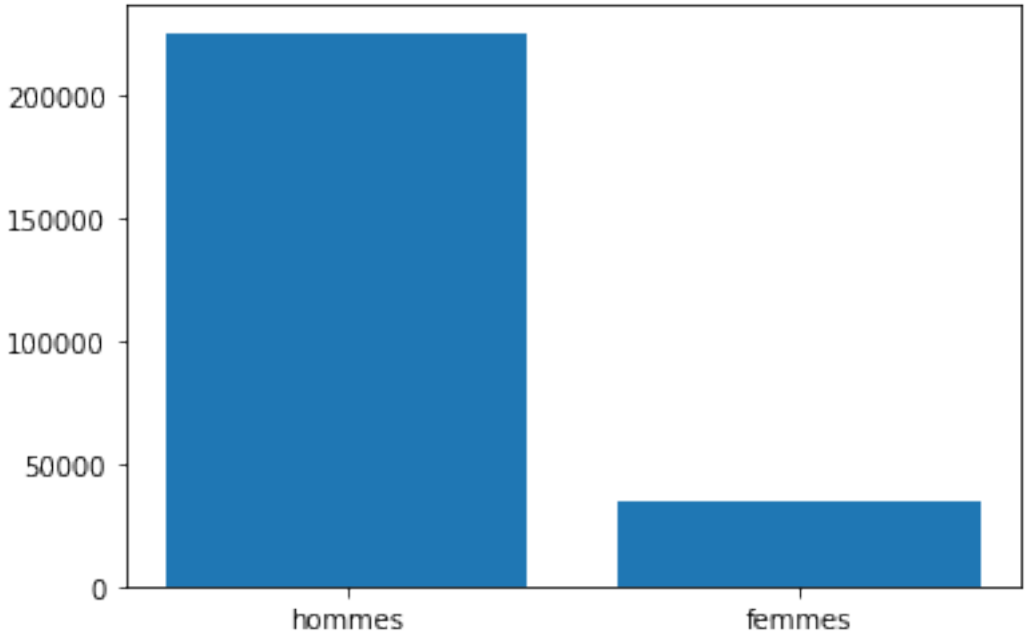

FIGURE 4.12 – Nombres absolus de *claims* selon le genre du/de la premier/ère auteur·ice de l'article

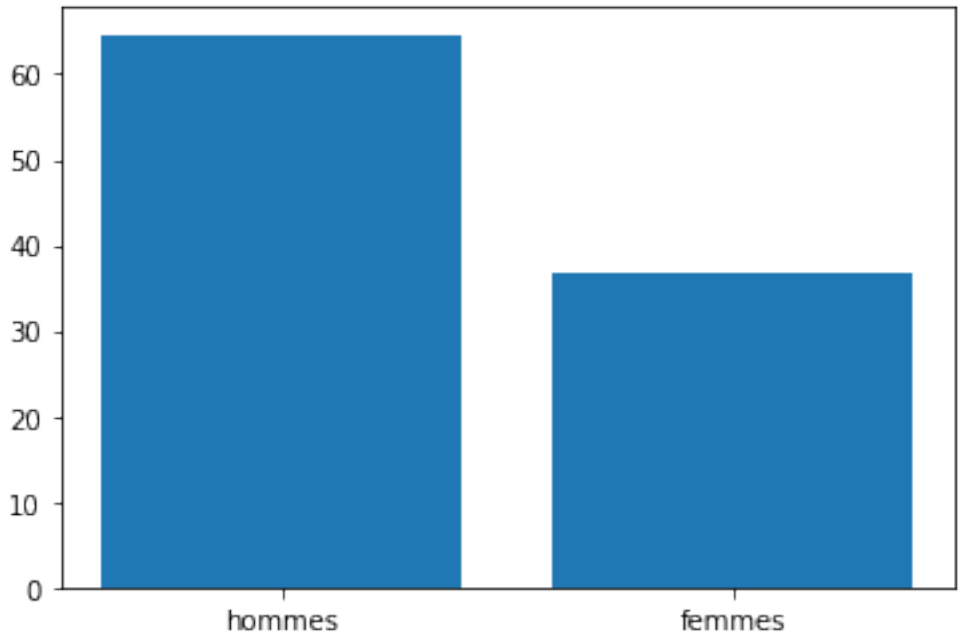

FIGURE 4.13 – Nombres relatifs de *claims* selon le genre du/de la premier/ère auteur·ice de l'article

la manière d'écrire les *claims* reste similaire. Les répartitions selon les forces et selon les parties sont très proches : les affirmations sont surtout de force 1, et presque autant de force 0 que de force 2, et elles sont réalisées surtout dans le corps d'article, puis en conclusion, puis presque équitablement en introduction et dans le résumé.

Pour conclure cette sous-section, il semblerait que le genre soit corrélé avec le nombre de *claims*, mais par sur le type ou l'endroit de ces derniers. Cette approche nous a également permis de voir plus généralement la proportion d'hommes et de femmes publié·es à ACL.

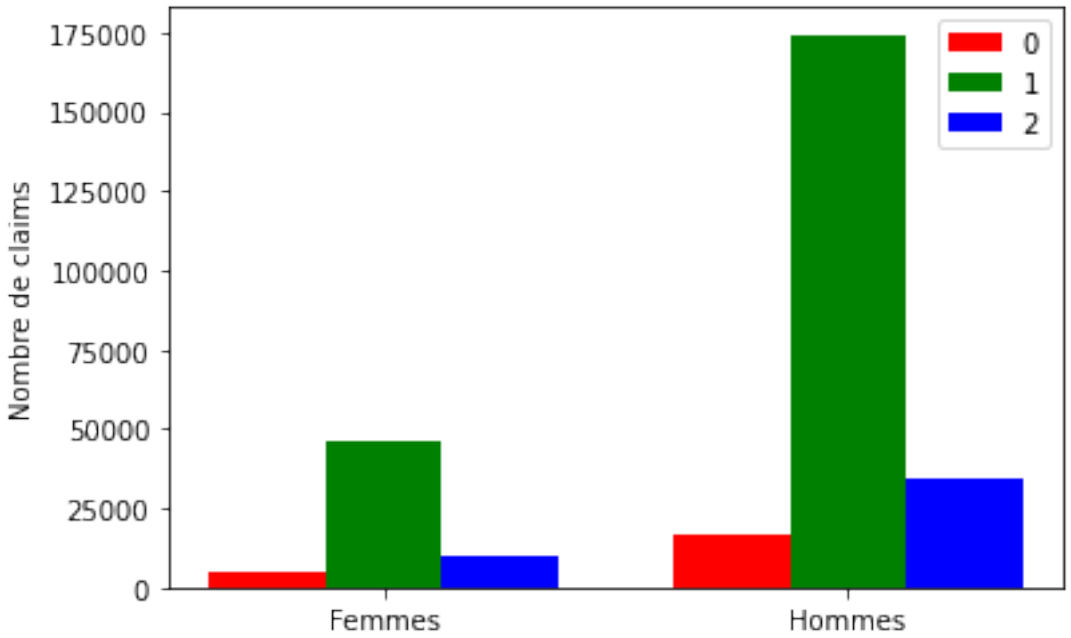

FIGURE 4.14 – Nombres absolus de *claims* par genre du/de la premier/ère auteur·ice selon la force

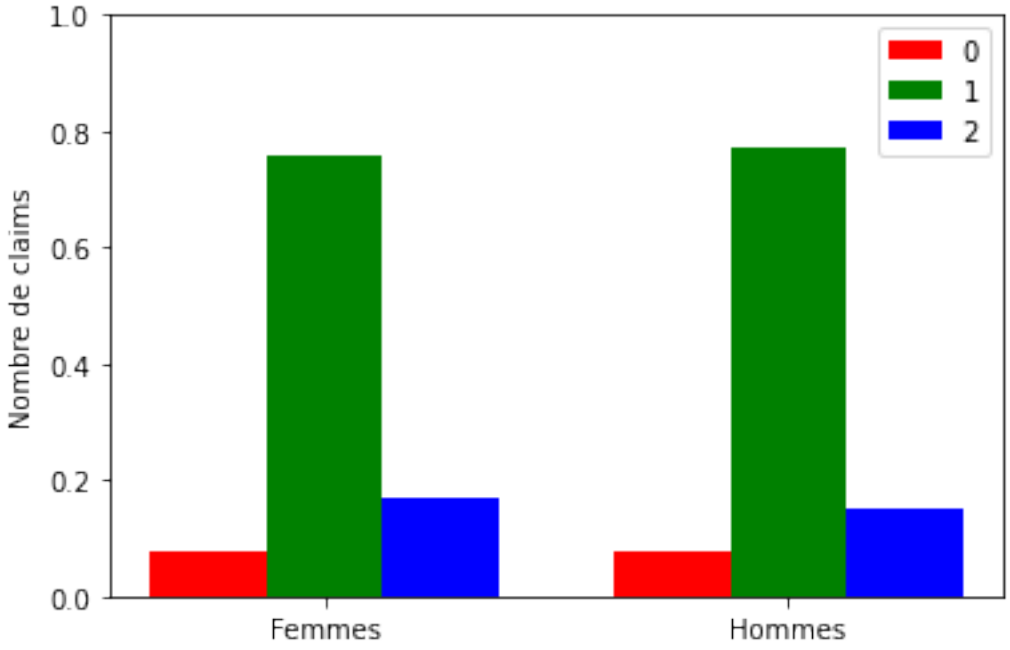

FIGURE 4.15 – Proportions de *claims* par genre du/de la premier/ère auteur·ice selon la force

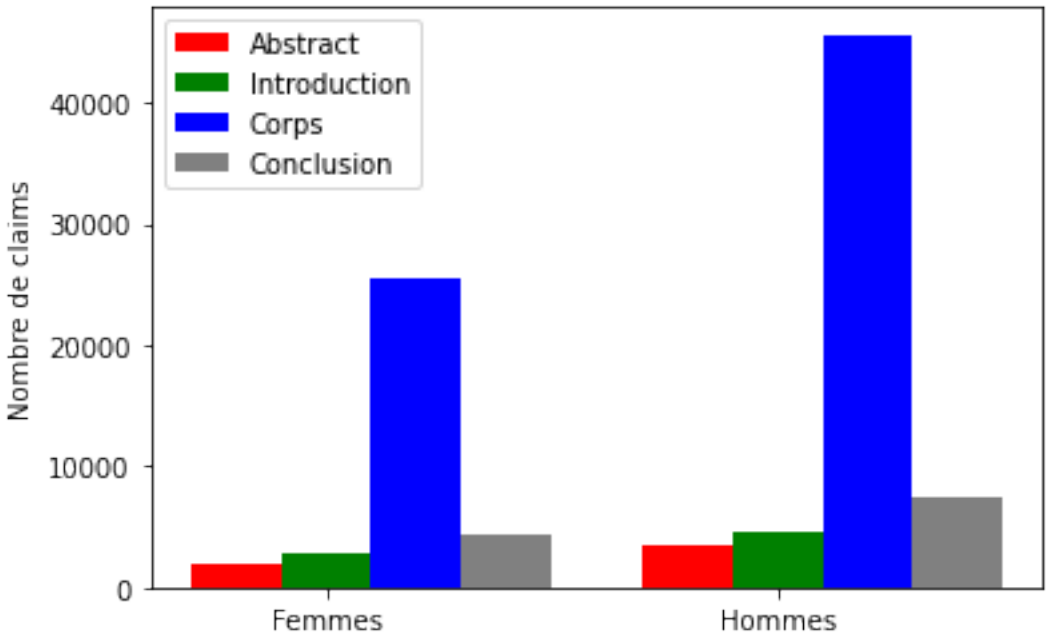

FIGURE 4.16 – Nombre de *claims* par force selon les parties et le genre du/de la premier/ère auteur·ice

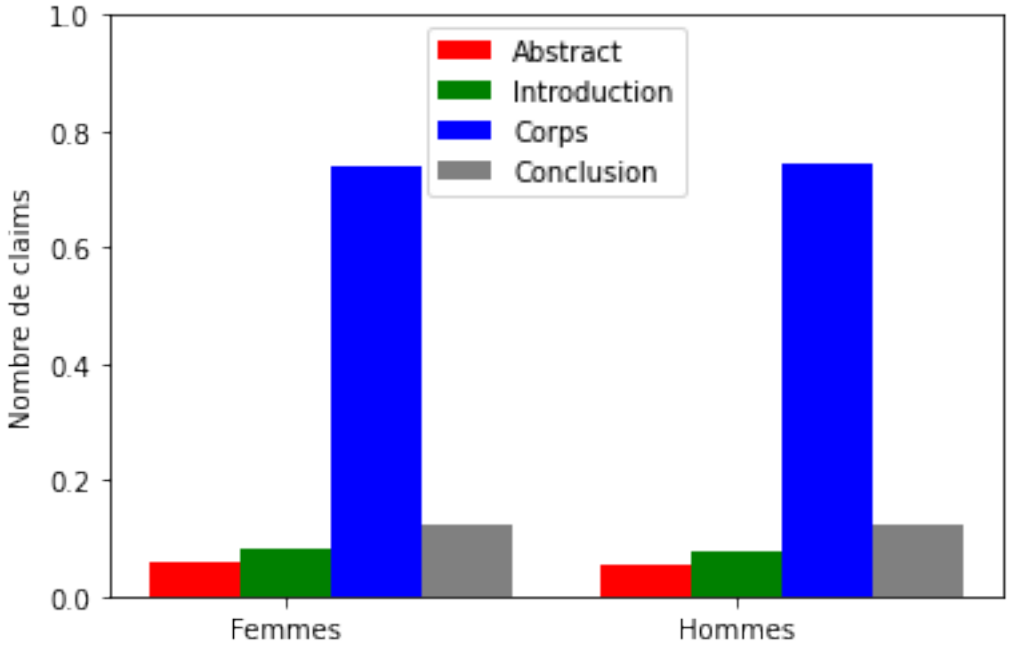

FIGURE 4.17 – Proportion de *claims* par force selon les parties et le genre du/de la premier/ère auteur·ice

4.2 Corrélation avec le continent

Nous réutilisons toutes les affirmations du corpus et les recoupons cette fois-ci avec des informations sur les continents des auteur·ices, en attribuant à un article le continent le plus affilié aux auteur·ices du groupe. Nous avons pu affilier 6 663 articles à leur continent majoritaire, cette sous-section couvre alors plus de 92 % du corpus total.

Ainsi, nous trouvons que 2 388 articles sont majoritairement affiliés à l'Amérique du Nord (soit quasiment 36 %), 2 070 à l'Europe (soit 31 %), 2 040 articles à l'Asie (soit plus de 30 % du corpus), 123 à l'Océanie (soit 1,8 %), 25 à l'Amérique du Sud (soit 0,37 %) et 19 à l'Afrique (soit 0,28 %). Nous représentons cela visuellement sur la figure 4.18 et tenons à faire remarquer qu'à elles seules, l'Asie, l'Europe et l'Amérique du Nord constituent plus de 97 % des affiliations du corpus total. Puis, nous réalisons le même graphique mais en prenant en compte le nombre de *claims* présents dans les articles affiliés aux continents correspondants (voir Figure 4.19).

Différences de quantité

Nous remarquons que ces deux figures sont proches et semblent être assez proportionnelles, ce qui n'est pas étonnant. En effet, un continent avec peu d'articles comptera également peu d'affirmations, à moins que ses articles en émettent énormément. Cependant, on peut déjà remarquer que l'écart entre l'Europe et l'Amérique du Nord se creuse sur la deuxième figure par rapport à la première, ce qui semble vouloir dire que les articles venant de ce dernier continent émettent proportionnellement plus de *claims* par article que ceux affiliés à l'Europe.

En considérant les nombres relatifs, c'est-à-dire en divisant le nombre d'affirmations du continent par son nombre d'articles, nous obtenons la figure 4.20. Grâce à celle-ci, nous constatons que ce sont les articles dont les auteur·ices sont majoritairement affilié·es à l'Amérique du Sud qui émettent le plus de *claims* (56,33 par article en moyenne), suivis de très près par les affiliations à l'Asie (56,32), puis par celles à l'Europe (51,41), à l'Océanie (47,1), à l'Afrique (46,31) et, finalement, à l'Amérique du Sud (avec une moyenne de 28,64 *claims* par article seulement, soit presque deux fois moins que pour l'Asie et l'Amérique du Nord). Nous calculons que la moyenne de toutes ces valeurs égale 47,68 affirmations par article (et que l'on compte en moyenne 468,5 phrases par article, soit un peu plus d'une phrase sur dix qui serait un *claim*).

On pourrait en déduire que les auteur·ices affilié·es à l'Asie, l'Amérique du Nord et l'Europe font preuve de plus de certitude, et donc potentiellement d'assurance. A l'inverse, les personnes affiliées à l'Amérique du Sud, et dans une moindre mesure, à l'Afrique, semblent plus humbles. Cela est probablement à corréler avec le nombre d'articles affiliés à ces continents. Il semblerait que plus un continent soit représenté, plus il émette de *claims*.

On pourrait également s'intéresser à l'utilisation plus globale de la modalité épistémique dans leur manière d'écrire en anglais, voire dans leur langue d'origine.

Nous nous intéressons ensuite brièvement aux proportions d'articles contenant au

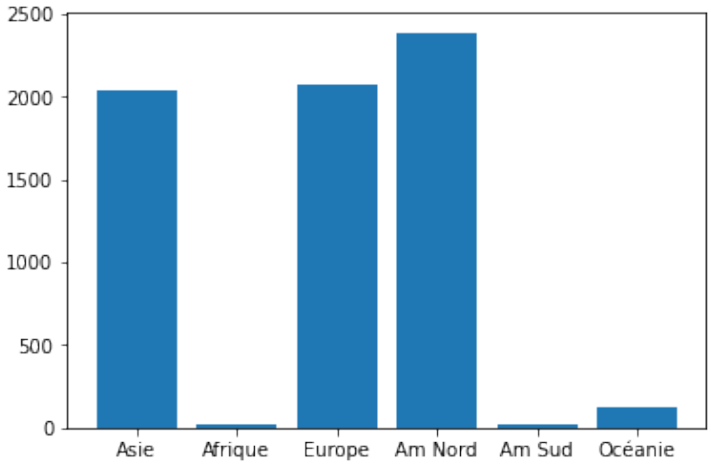

FIGURE 4.18 – Nombre d'articles selon le continent majoritairement affilié

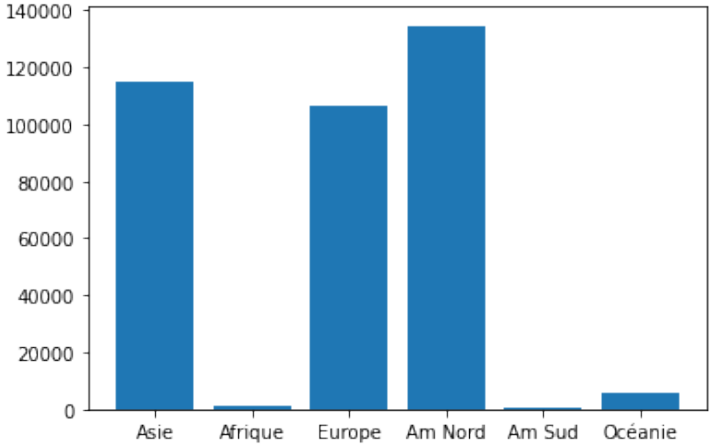

FIGURE 4.19 – Nombre de *claims* selon le continent majoritairement affilié

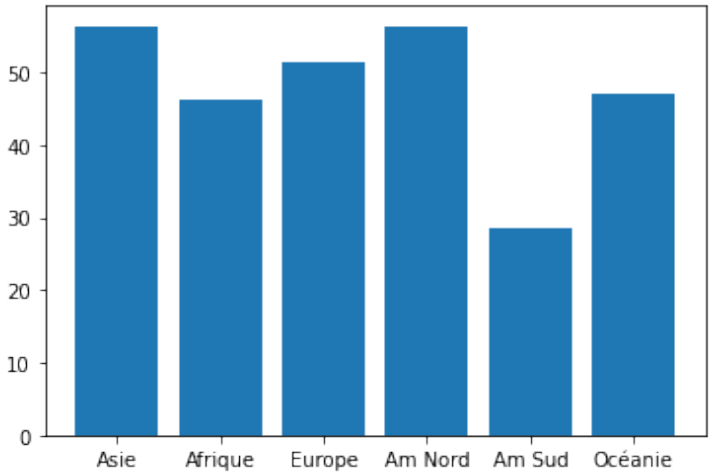

FIGURE 4.20 – Nombre moyen de *claims* par article selon le continent majoritairement affilié

moins un *claim* pour chaque continent. Nous trouvons que toutes ces proportions sont comprises entre 96,1 % (pour l'Amérique du Sud) et 100 % (pour l'Afrique, en gardant à l'esprit que nous avons un total de 19 articles seulement). Toutes les autres proportions sont comprises entre 99,1 % et 99,5 %, soit juste un peu plus que la moyenne globale de 99 % évoquée dans la section 3.4.1.

Différences de degrés

Nous reprenons ensuite les idées de la sous-section sur le genre et réalisons les graphiques prenant en compte la répartition des *claims* selon leur force d'une part (voir Figure 4.21), et selon leur localisation dans l'article d'autre part (voir Figure 4.22). Nous ne faisons figurer ici que les graphiques avec les nombres relatifs.

Au premier abord, la figure 4.21 semble refléter des répartitions très proches, voire identiques. En effet, le degré 1 est toujours largement majoritaire et avec des proportions très proches, bien que l'Europe, l'Amérique du Nord et l'Océanie aient les valeurs les plus élevées avec des proportions allant de 77 % à 79 %.

Cependant, si l'on s'y intéresse plus en détails, on remarque certaines différences, notamment dans la répartition entre degrés 0 et 2. Par exemple, l'Asie et l'Amérique du Sud se distinguent avec respectivement 17 et 20 % de *claims* de degré 2 tandis que cette proportion de *claims* certains est toujours comprise entre 14 et 15 % pour les autres continents. Nous pouvons toutefois noter que cette plus grande proportion pour l'Asie et l'Amérique du Sud est compensée par une proportion plus basse de *claims* de degré 1. À l'inverse, la proportion de *claims* incertains reste toujours très proche pour les autres continents, elle est toujours comprise entre 6,9 et 7,8 %, à l'exception de l'Afrique. Dans ce continent, la proportion de *claims* incertains atteint les 8,9 %.

Nous pouvons alors remarquer que les deux continents les moins représentés dans notre corpus adoptent deux attitudes opposées : l'Afrique a la plus grande proportion de *claims* incertains tandis que l'Amérique du Sud a la plus grande proportion de *claims* certains. On peut toutefois se demander s'il s'agit d'attitudes adoptées en réaction à leur sous-représentation. Par exemple, on pourrait penser que les auteur·ices d'Afrique se sentent moins légitimes et ont moins confiance en leurs résultats tandis que les auteur·ices d'Amérique du Sud essayent d'écrire avec plus de certitude afin de s'efforcer de se faire une place dans la communauté scientifique. Toutefois, d'autres hypothèses pourraient également être présentées. Nous pourrions imaginer que la maîtrise de la langue anglaise, dans laquelle les auteur·ices doivent rédiger leurs articles, est différente, et que la modalité épistémique de leur langue maternelle a un impact sur leur manière d'utiliser la modalité épistémique en anglais.

La figure 4.22 permet également de dégager de grandes tendances : les *claims* sont toujours majoritairement présents dans le corps puis, dans une bien moindre mesure, dans la conclusion, puis l'introduction, puis le résumé (à l'exception de l'Afrique, où l'ordre est : corps, résumé, introduction, conclusion). L'autre continent qui semble sortir du lot est l'Amérique du Sud, où la proportion d'affirmations en conclusion est plus élevée qu'ailleurs, et en conséquence la proportion dans les corps d'articles un peu moins importante. En Amérique du Sud, la conclusion semble donc être la partie où l'on peut exprimer plus clairement ses certitudes. Cette remarque parait également pertinente pour les autres continents (si l'on met de côté la prévalence de la catégorie *corps*, qui s'explique par la longueur bien plus importante de cette partie). On peut ajouter que le résumé

semble être la partie où l'on émet le moins de *claims*. Cela s'avère rassurant si l'on pense aux remarques de Koroleva [2017] mentionnées dans la section 1.1, qui mettait en lumière les conséquences néfastes liées à la présence de *claims* trop prétentieux dans les résumés.

Représentation des différents continents dans le corpus

Nous nous arrêtons un instant pour faire d'autres remarques plus générales sur notre corpus. Comme pour le genre, nous réalisons un graphique, 4.23 , représentant l'évolution diachronique de la présence des différents continents. Cependant, celui-ci n'étant pas très lisible, nous en réalisons deux autres, un allant de 1979 à 2000 et l'autre de 2000 à 2020, que nous joignons aux Annexes 6.5.

Nous pouvons distinguer la grande importance de l'Amérique du Nord, l'Europe et l'Asie, qui sont les courbes qui se détachent des autres et qui fluctuent (et augmentent) le plus. L'Amérique du Nord surplombe les autres dès 1979, mais se voit devancer par l'Asie et/ou l'Europe à partir de 1998. Par la suite, les trois continents alternent la première position selon les années, jusqu'à ce que l'Amérique du Nord regagne en puissance et repasse devant en 2016, à l'inverse de l'Europe qui semble connaître un certain déclin. Par ailleurs, l'Afrique, l'Amérique du Sud et l'Océanie restent très peu représentées, bien que l'on observe de légers sursauts certaines années pour l'Océanie.

Nombre d'autrices par continent

Nous recoupons ensuite les différentes données que nous avons collectées sur le genre et les continents afin de représenter les proportions d'articles écrits majoritairement par des hommes, des femmes, ou une égalité d'hommes et de femmes selon les continents. Nous en réalisons un graphique avec des pourcentages (voir Figure 4.24) et mettons celui avec les nombres absolus en Annexes.

Nous remarquons que, mis à part en Afrique où l'égalité entre les articles écrits par une majorité d'hommes et ceux écrits par une majorité de femmes est proche, et même où il y a plus d'articles écrits par une majorité de femmes (30 % pour les hommes et 40 % pour les femmes), les autres continents ont une très large majorité d'articles écrits majoritairement par des hommes (entre 75 et 85 %). La barre des 80 % est atteinte par l'Amérique du Sud et l'Océanie. Cela pourrait être à nuancer par le faible nombre absolu d'articles affiliés majoritairement à ces deux continents, surtout pour l'Amérique du Sud. L'Océanie se démarque également par le fait que ce soit le seul continent où la catégorie égalité dépasse celle de la majorité de femmes (qui est la plus faible du lot, avec seulement 4 % des articles), tandis qu'à l'inverse, l'Amérique du Sud n'a aucun article écrit par autant d'hommes que de femmes.

Finalement, on peut remarquer que les trois continents les plus présents dans le domaine du TAL, l'Asie, l'Europe et l'Amérique du Nord, ont des répartitions très similaires, avec entre 75 et 78 % d'articles écrits majoritairement par des hommes, 11 ou 12 % majoritairement par des femmes et entre 9 et 11 % par autant d'hommes que de femmes.

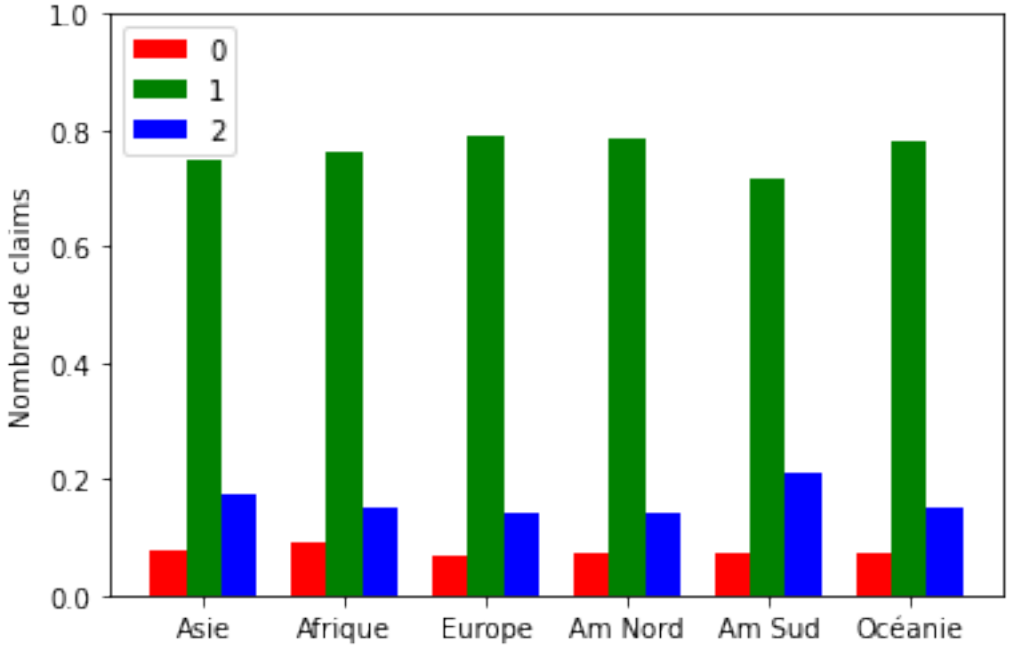

FIGURE 4.21 – Répartition des *claims* selon leur force et le continent d'affiliation majoritaire

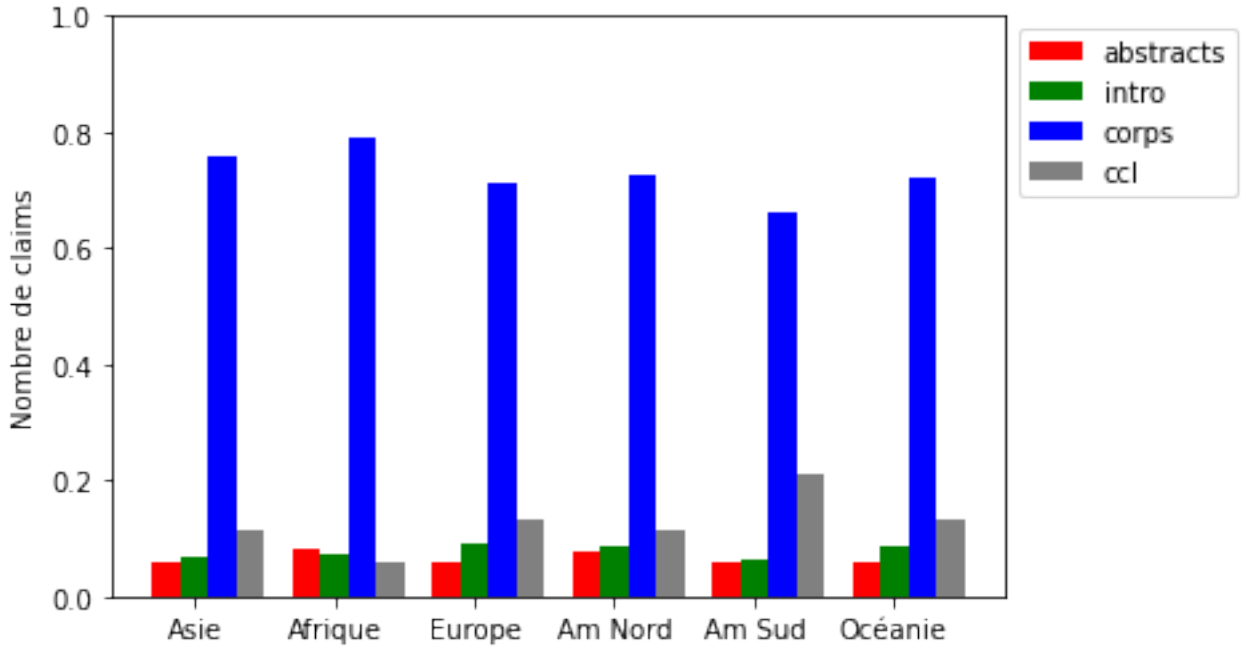

FIGURE 4.22 – Répartition des *claims* selon leur partie et le continent d'affiliation majoritaire

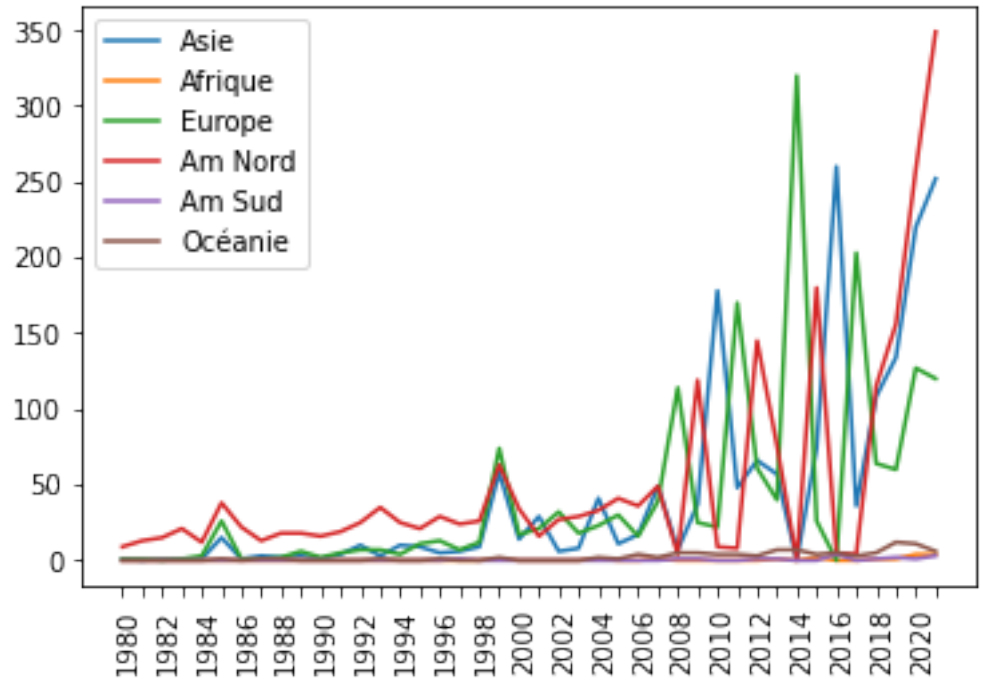

FIGURE 4.23 – Evolution diachronique du nombre d'articles majoritairement affilié aux différents continents

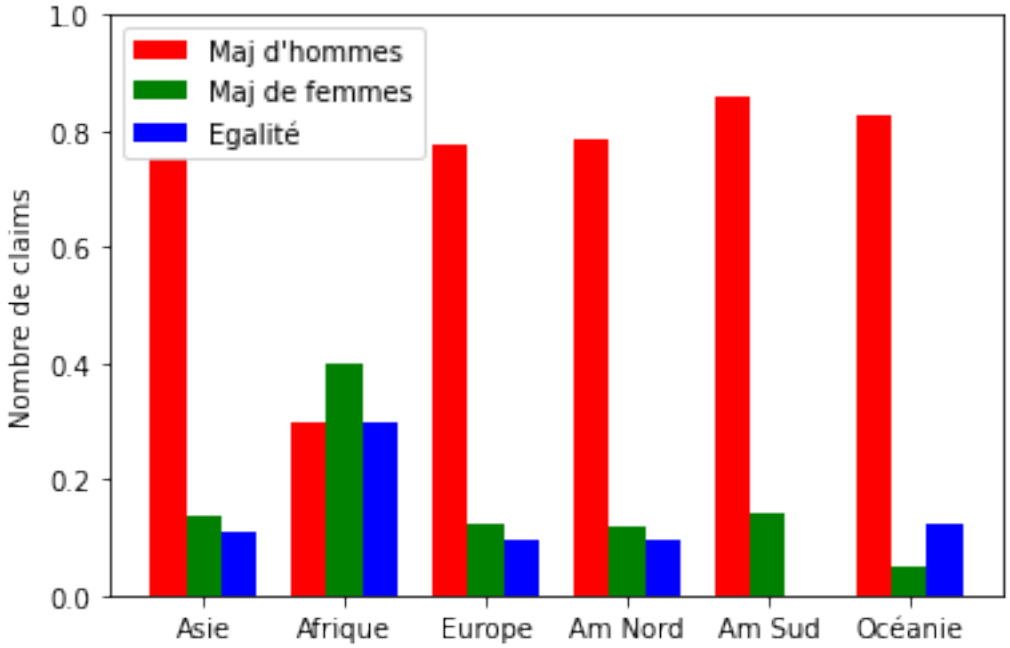

FIGURE 4.24 – Répartition des groupes d'auteur·ices selon les continents

4.3 Corrélation avec des affiliations à des institutions de prestige

Suite à notre travail de détection d'institutions de prestige, à savoir ici les GAFAM et les 50 premières universités du classement de Shanghai, nous trouvons que parmi notre corpus de 6 372 articles, 2 401 sont affiliés à des institutions de notre liste (soit 37 %). Parmi eux, 2 394 contiennent au moins un *claim* (soit 99,7 %).

Bien que les articles contenant des affiliations à des institutions ne représentent que 37 % de notre corpus, on constate que ce sont 41 % des *claims* présents dans tout le corpus qui sont en réalité présents dans des articles affiliés à des institutions : on atteint en effet une moyenne de 66,7 *claims* (ou indicateurs de *claims*) par article), contre une moyenne de 58 par article pour tout le corpus (voir Figure 4.25).

Nous illustrons cela par les figures suivantes 4.29 (attention, chaque graphique a sa propre échelle et présente un phénomène différent, ils ne sont donc pas comparables les uns par rapport aux autres).

Comme dans les sous-sections précédentes, nous nous intéressons à la répartition des affirmations selon leur force et selon leur localisation dans l'article. Nous pouvons en tirer des conclusions similaires à celles faites dans la sous-section sur le genre : la différence principale semble être quantitative et non pas qualitative, les affirmations issues d'articles affiliés à des institutions sont bien plus nombreuses mais situées aux mêmes endroits et avec les mêmes degrés que celles qui ne sont affiliées à aucune institution.

Afin de rendre mieux compte de cette différence quantitative, nous laissons les figures avec les nombres absolus, en rappelant que les articles affiliés à des institutions ne représentent que 37 % des articles totaux du corpus mais que leurs valeurs absolues sont en réalité très proches de celles des articles non affiliés.

La différence de degrés n'est en effet pas significative (voir Figure 4.31), on retrouve, comme dans les sections précédentes, une large majorité de *claims* de degré 1, et une très légère supériorité des *claims* de degré 2 par rapport à ceux de degré 0.

Ainsi, nous pouvons en conclure que les auteur·ices qui ont une affiliation à une insti-

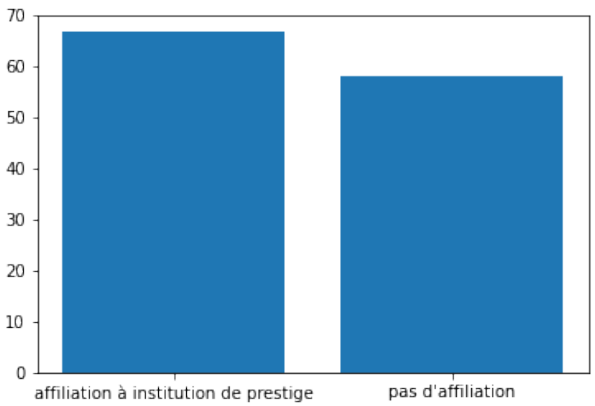

FIGURE 4.25 – Nombre moyen de *claims* par article selon les affiliations à des institutions de prestige

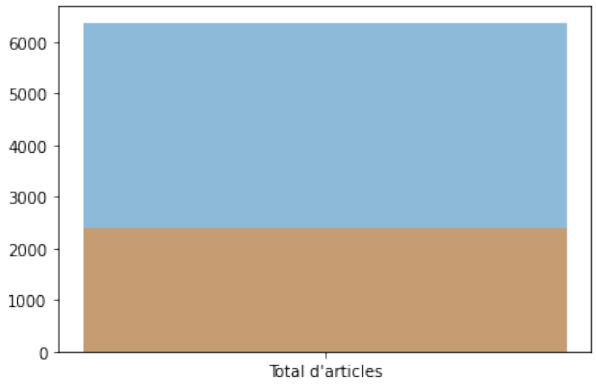

FIGURE 4.26 – Proportion d'articles contenant des institutions de prestige (en marron) parmi tout le corpus (en bleu)

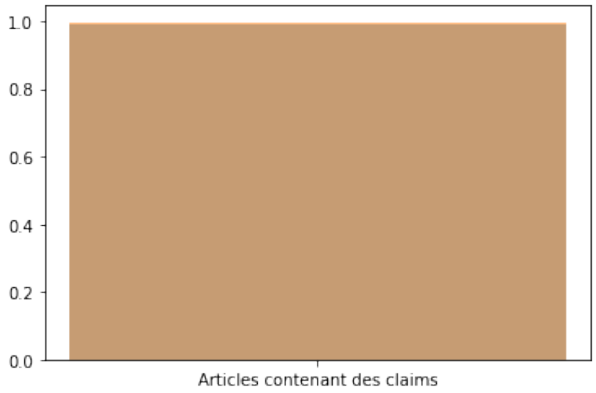

FIGURE 4.27 – Proportion d'articles contenant des *claims* et des institutions de prestige (en marron) parmi tous les articles contenant des claims (en bleu)

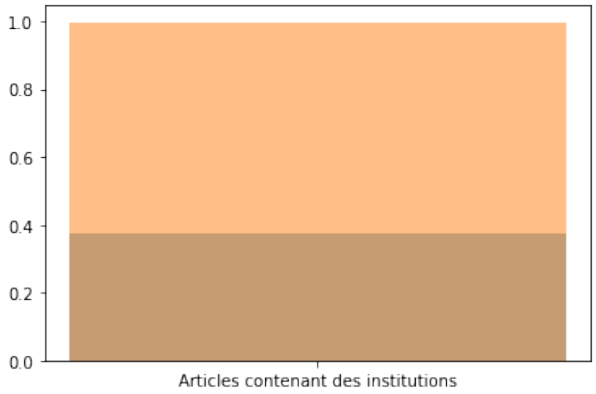

FIGURE 4.28 – Proportion d'articles contenant des *claims* et des institutions de prestige (en marron) parmi tous les articles contenant des institutions (en marron clair)

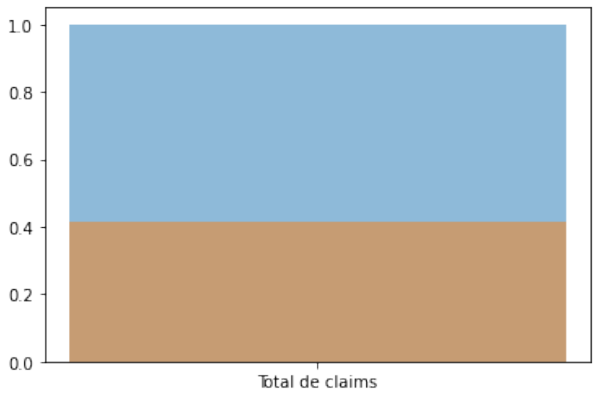

FIGURE 4.29 – Proportion de *claims* présents dans des articles affiliés à des institutions de prestige parmi tous les *claims* du corpus

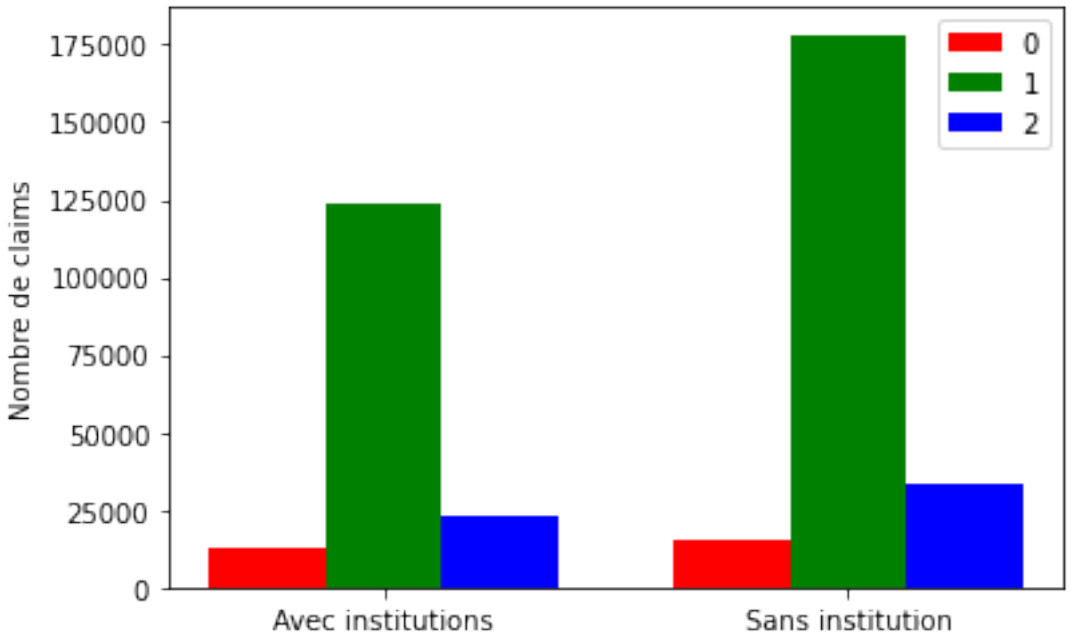

FIGURE 4.30 – Nombres absolus de *claims* selon leur degré et leur potentielle affiliation à des institutions

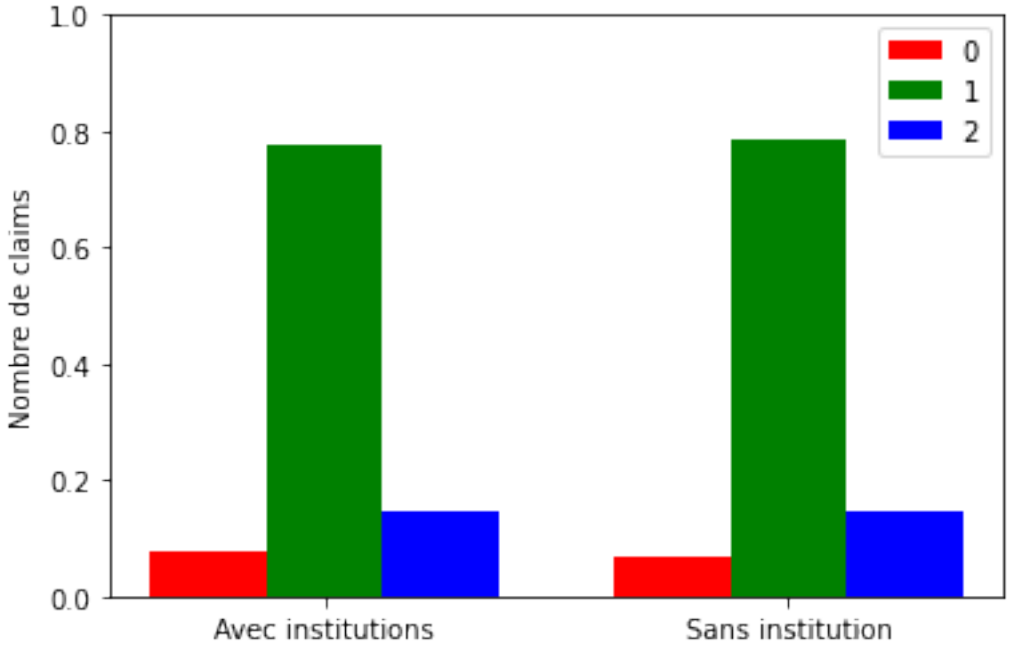

FIGURE 4.31 – Proportions de *claims* selon leur degré et leur potentielle affiliation à des institutions

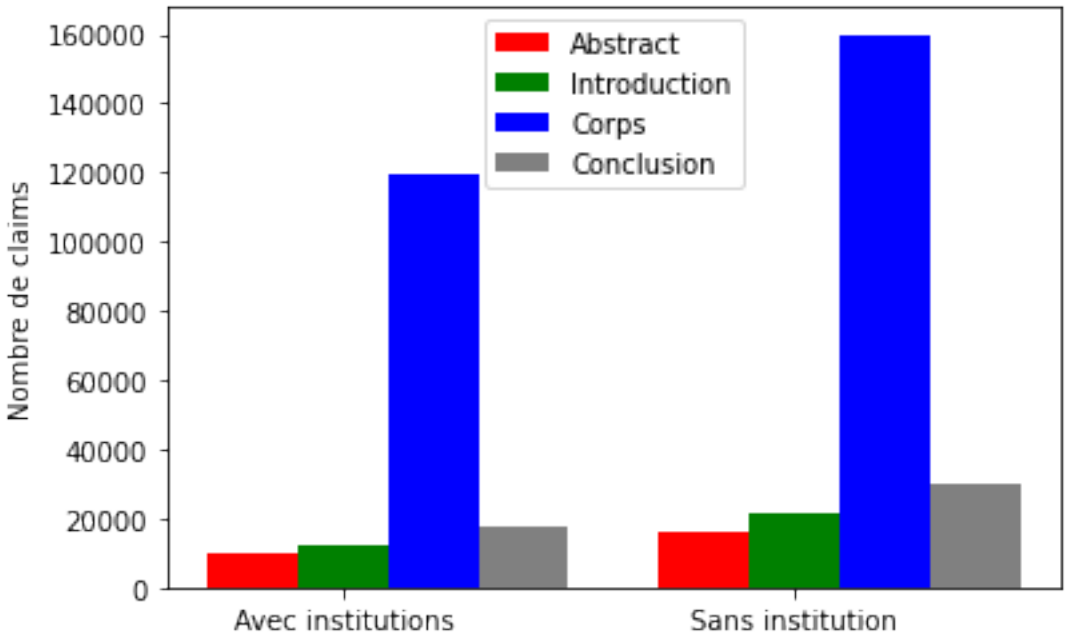

FIGURE 4.32 – Nombres absolus de *claims* selon leur localisation et leur potentielle affiliation à des institutions

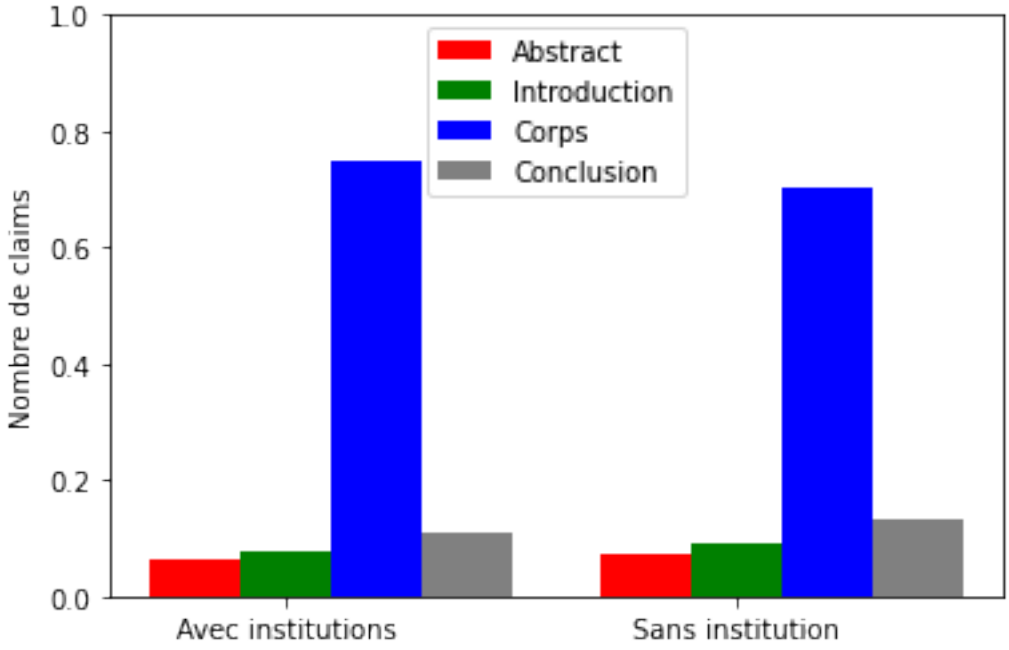

FIGURE 4.33 – Proportion de *claims* selon leur localisation et leur potentielle affiliation à des institutions

tution de prestige expriment plus de certitude. A l'inverse du syndrome de l'imposteur qui impacte les femmes, et peut-être également les personnes affiliées à des continents moins représentés dans le domaine du TAL, nous pourrions penser que leur affiliation leur offre un sentiment fort de légitimité.

Nous pourrions ajouter à cela un facteur économique. En effet, il est très probable que les personnes affiliées à de grandes institutions aient plus de moyens techniques et financiers à leurs dispositions. Ces meilleures conditions de recherche peuvent avoir une influence sur leur travail et l'améliorer. De ce fait, l'obtention de résultats satisfaisants est facilitée, tout comme l'expression de *claims*, qui seraient dans ce cas légitimes.

Chapitre 5

Conclusion

Nous avons proposé un état de l'art abordant des domaines variés afin de bien définir les notions linguistiques et informatiques utilisées pendant le reste de l'étude, nous avons ensuite extrait plus de 300 000 *claims* dans un corpus de plus de 6 000 articles en anglais d'ACL, la conférence internationale la plus renommée du domaine du TAL. Il s'agit du plus grand corpus utilisé dans une étude sur les *claims* scientifiques et le premier portant sur le TAL. À partir de ceux-ci, nous avons déjà pu tirer quelques conclusions sur la nature et la construction des *claims*.

Nous avons ensuite analysé ces articles avec des algorithmes de *clustering* afin de les catégoriser selon leur degré de certitude, ou, autrement dit, la force de leur modalité épistémique. Nos métriques d'évaluation ont révélé que la meilleure combinaison possible est d'utiliser trois *clusters* sur la version lemmatisée du corpus en conservant les mots grammaticaux. Nous avons ainsi pu remarquer que la majorité des affirmations sont moyennement certaines et qu'on les retrouve surtout dans les corps d'articles et les résumés. Cependant, selon la partie et la force étudiée, les résultats changent. Nous avons également constaté que les auteur·ices ont tendance à émettre des *claims* de plus en plus certains au fil de l'article, ou à rester à un degré moyennement certain tout du long. Par ailleurs, notre approche diachronique nous a permis de remarquer que l'évolution du nombre de *claims* par article n'est pas aussi linéaire que ce que l'on aurait pu penser, que la proportion de *claims* moyennement certains tend à décroître au fil du temps et que les *claims* sont de plus en plus présents dans les corps d'articles, au détriment des autres parties.

Nous avons finalement recoupé ces données avec d'autres informations liées aux auteur·ices : genre, continent et institution. Cela nous a permis de remarquer qu'une majorité d'articles est encore écrite par des hommes et/ou des personnes affiliées à l'Amérique du Nord, l'Asie ou l'Europe, et que, dans plus d'un tiers des cas, une grande institution y est rattachée. La corrélation avec ces facteurs sur les *claims* est surtout quantitatif. Les catégories les plus représentées sont celles qui émettent le plus de *claims*, bien qu'ils soient globalement de même force et dans les mêmes parties d'articles que ceux des autres populations. Nous pouvons alors en conclure que c'est la quantité de *claims* qui change plus que leur nature, et qu'il existe un profil type de personne qui en écrit énormément : les hommes venant d'Asie, d'Amérique du Nord ou d'Europe et étant affiliés à une institution de prestige. Leurs travaux pourraient alors être plus mis en avant et cités, alors que leur qualité ne serait pas pour autant assurée.

Cela peut nous permettre d'en tirer des conclusions plus larges. En effet, le fait que ce soient les populations les plus représentées (et privilégiées) qui émettent le plus de *claims* a diverses implications : on aura sûrement tendance à encore plus valoriser leur travail, qui semble porter plus de fruits, que celui des autres, qui, peut-être à cause d'un sentiment d'illégitimité, ont tendance à plus se surveiller voire même se censurer quand iels ne sont pas totalement certain·s de leurs remarques. Nous pouvons également penser à

d'autres exemples plus concrets : les langues les moins dotées sont majoritairement celles parlées dans les continents les moins représentés à ACL, alors que ce sont les auteur·ices qui y sont affilié·es qui sont les plus à mêmes de travailler dessus et de permettre de faire avancer la recherche, et donc la mise en place d'outils performants [Ducel et al., 2022].

Nous aimerions reproduire ce travail sur un corpus d'articles n'ayant pas été acceptés à ACL afin de voir si les tendances sont les mêmes, ou si leur rejet pourrait être expliqué par les *claims* qu'ils comportent.

On peut par exemple penser que les articles qui comportent beaucoup de *claims* très certains ont été rejetés car les résultats ne semblaient pas réalistes, ou, à l'inverse, penser aux cas des articles avec trop peu de *claims* et/ou avec une majorité de *claims* incertains, qui laissent penser que l'étude n'est pas assez fructueuse pour être publiée. Cela permettrait aussi de mener les analyses liées aux autres facteurs sociologiques et de voir si l'on trouve une sur-représentation de certaines catégories dans ce corpus d'articles rejetés. Ainsi, nous pourrions en conclure que la trop grande utilisation d'*hedging* et l'incertitude peuvent être néfastes à certaines catégories de population.

Nous souhaiterions également utiliser des corpus d'articles issus d'autres conférences pour étudier leurs tendances. Cette expérience pourrait être particulièrement intéressante dans le cas de conférences nationales comme TALN puisque les conclusions et comparaisons avec la présente étude permettraient de s'intéresser aux spécificités de populations plus restreintes. Nous pourrions aussi envisager de nous intéresser à des articles d'autres domaines scientifiques, voire à des écrits de genres complètement différents ; les enjeux éthiques seraient alors bien différents.

Si nous restons dans la continuation de cette étude et de ce corpus en particulier, nous pourrions penser à améliorer le découpage des articles selon leur structure, notamment pour affiner la partie « corps d'articles ». Une autre limite de notre étude est l'utilisation d'une liste d'indices qui, bien que basée sur l'état de l'art, demeure arbitraire. Toutefois, nos codes sont librement disponibles [1], cette liste peut donc être modifiée facilement et les expériences relancées avec de nouveaux indices.

L'une des limites de notre étude réside dans le fait qu'elle soit totalement automatique et non-supervisée. Nous continuons actuellement nos expériences et commençons un processus d'annotation manuelle. Cela nous permettra de réaliser une évaluation plus extrinsèque de l'extraction et la classification des *claims*, mais aussi de corriger le bruit que nous pourrions détecter pour améliorer notre système.

Pour aller plus loin, nous pourrions imaginer la mise en place d'un système de détection des *claims* forts, de degré 0 ou 2, qui pourrait avertir l'auteur·ice ou le lectorat afin qu'une attention particulière soit portée sur ces phrases. Les *claims* de degré 0 sont inclus dans cette remarque, car, comme suggéré par Bowman [2022], sous-estimer ses résultats peut également poser problème. Selon lui, affirmer qu'un système est moins puissant qu'il ne l'est réellement empêche d'anticiper ses effets négatifs et de chercher à les contrer. Le but n'est pas d'essayer d'éliminer ces degrés de certitude, mais de faire en sorte que l'on vérifie si ces degrés sont vraiment adaptés.

Cela pousserait notamment à s'intéresser au phénomène d'embellissement, c'est-à-dire

1. https://github.com/FannyDucel/MemoireM1Claims

à la cohérence entre les *claims* et l'intensité des résultats effectivement obtenus. Cela implique également de vérifier que ces chiffres sont réellement significatifs, que les interprétations qui en découlent sont plausibles et que les métriques d'évaluation choisies sont les plus pertinentes. Ce dernier point reste particulièrement difficile à aborder, notamment du fait que la significativité des résultats dépend de la popularité de la tâche : une augmentation de X points de telle métrique n'a pas le même impact si elle est liée à une tâche relativement aisée et déjà réalisée moult fois que si elle est liée à une tâche difficile réalisée sur une langue peu dotée. Cela pourrait aussi servir à détecter les procédés révélant des affirmations inappropriées selon la typologie de Koroleva [2017] (voir Section 1.1) : effets négatifs omis, contexte de l'étude imprécis, utilisation abusive de *spin*, affirmation sur des résultats non significatifs, et autres extrapolations inappropriées.

Nous concluons en rappelant l'importance et les enjeux éthiques qui se cachent derrière les *claims* et leur modalité épistémique, et en réaffirmant que la fausse modestie/une trop grande incertitude tout comme un manque d'humilité ou des extrapolations peuvent être nuisibles au domaine du TAL et à la diffusion des résultats de la science. Cela est d'autant plus vrai à l'heure des résumés automatiques, friands de *claims*, au même titre que la mémoire humaine, qui peut facilement retenir des informations faussées ou biaisées en se souvenant uniquement des phrases les plus fortement marquées du point de vue de la modalité épistémique.

Chapitre 6

Annexes

6.1 Liste des institutions utilisées en 2.4.3

Google, Amazon, Facebook, Apple, Microsoft, Harvard University, Stanford University, University of Cambridge, Massachusetts Institute of Technology, MIT, University of California, Princeton University, University of Oxford, Columbia University, California Institute of Technology, Caltech, University of Chicago, Yale University, Cornell University, Université Paris-Saclay, University of California, University of Pennsylvania, Johns Hopkins University, University College London, University of Washington, Swiss Federal Institute of Technology Zurich, ETHZ, University of Toronto, Washington University, University of Tokyo, Imperial College London, University of Michigan, New York University, Tsinghua University, University of North Carolina, University of Copenhagen, University of Wisconsin-Madison, Duke University, University of Melbourne, Northwestern University, University of Manchester, Sorbonne University, Sorbonne Université, Kyoto University, University of Edinburgh, PSL Research University Paris,Université PSL,Université Paris Sciences Lettres, University of Minnesota, University of Texas at Austin, University of British Columbia, Karolinska Institute, Rockefeller University, Peking University, University of Colorado at Boulder, King's College London, University of München, University of Texas Southwestern Medical Center at Dallas, Utrecht University.

6.2 Tableaux des tokens présents dans la liste d'indices, précédant ou suivant un indice verbal dans une allégation (3.1)

6.2 Tableaux des tokens présents dans la liste d'indices, précédant ou suivant un indice verbal dans une allégation (3.1)

words	analyze	appear	assume	best	can	certain	claim	clearly	comment	conclude
analyze										
appear					x					
assume										
best										
can	x	x	x	x	x		x	x	x	x
certain										
claim										
clearly										
comment										
conclude										
confirm										
could										
demonstrate								x		
discover										
evidence										
examine										
explain										
find					x					
greatly										
guarantee										
indeed										
indicate					x			x		
infer										
lead					x					
may		x	x						x	
might		x								
most										
must		x								
note										
observe					x			x		
possible										
possibly										
predict					x					
prove										
report										
reveal										
seem										
should		x								
show					x			x		
significant										
significantly										
state				x	x	x				
suggest					x					
support					x		x			
think										

words	confirm	could	demonstrate	discover	evidence	examine	explain	find	greatly
analyze									
appear									
assume									
best									
can	x			x	x	x	x	x	x
certain									
claim									
clearly									
comment									
conclude									
confirm									
could							x	x	
demonstrate									
discover									
evidence									
examine									
explain									
find		x			x				
greatly									
guarantee									
indeed									
indicate		x							
infer									
lead		x							
may							x	x	
might								x	
most									
must								x	
note									
observe		x							
possible									
possibly									
predict		x							
prove									
report									
reveal									
seem									
should								x	
show					x				
significant									
significantly									
state			x						
suggest					x			x	
support					x			x	
think									

6.2 Tableaux des tokens présents dans la liste d'indices, précédant ou suivant un indice verbal dans une allégation (3.1)

words	guarantee	indeed	indicate	infer	lead	may	might	most	must	note	observe
analyze											
appear						x	x		x		
assume											
best											
can	x	x	x	x	x						x
certain											
claim											
clearly											
comment											
conclude											
confirm											
could			x		x						x
demonstrate											
discover											
evidence											
examine											
explain											
find						x	x		x		
greatly											
guarantee											
indeed											
indicate						x	x				
infer											
lead						x	x				
may			x		x	x					
might			x		x						
most											
must											
note											
observe											
possible											
possibly											
predict									x		
prove											
report											
reveal											
seem											
should					x					x	
show											
significant											
significantly											
state						x		x			x
suggest						x					
support											
think											

words	possible	possibly	predict	prove	report	reveal	seem
analyze							
appear							
assume							
best							
can		x	x	x	x	x	
certain							
claim							
clearly							
comment							
conclude							
confirm							
could		x	x	x			
demonstrate							
discover							
evidence							
examine							
explain							
find							
greatly							
guarantee							
indeed							
indicate							
infer							
lead							
may				x			x
might							x
most							
must			x				
note							
observe							
possible							
possibly							
predict							
prove							
report							
reveal							
seem							
should			x				
show							
significant							
significantly							
state	x				x		
suggest							
support							
think							

words	should	show	significant	significantly	state	suggest	support	think
analyze								
appear	x							
assume								
best								
can		x		x	x	x	x	x
certain								
claim								
clearly								
comment								
conclude								
confirm								
could								
demonstrate					x			
discover								
evidence								
examine								
explain								
find	x		x			x	x	
greatly								
guarantee								
indeed								
indicate								
infer								
lead	x							
may					x	x		x
might								x
most								
must								
note								
observe			x		x			
possible								
possibly								
predict	x							
prove								
report								
reveal								
seem								
should								
show			x		x			
significant								
significantly								
state		x			x			
suggest						x		
support							x	
think								

6.3 Graphiques de visualisation des méthodes Elbow

Attention, les graphiques suivants ne sont pas à la même échelle, la valeur maximale sur l'axe des ordonnées n'est pas la même sur chacune des figures. Toutefois, nous avons décidé de procéder ainsi afin que la courbe soit bien visible (sans effet de zoom arrière) et que le coude soit plus facilement détectable.

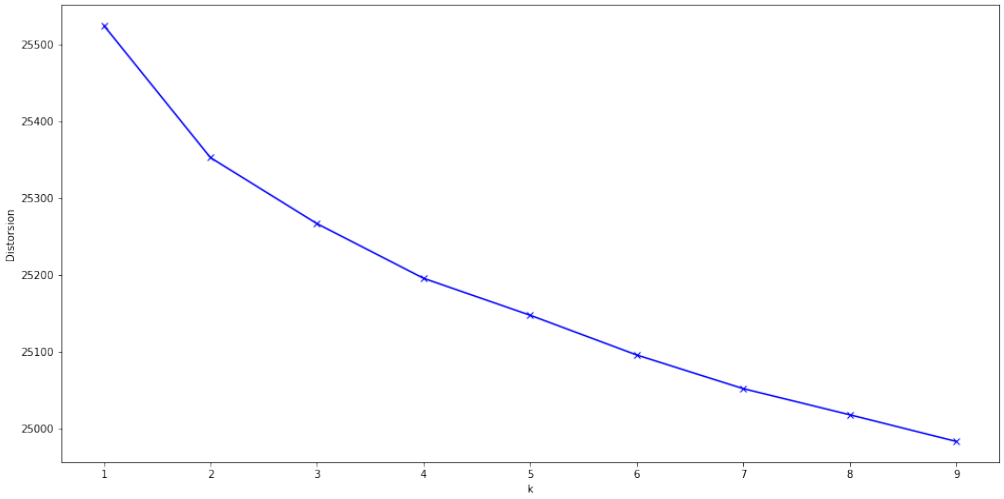

FIGURE 6.1 – Résultat de la méthode Elbow sur les *claims* des abstracts

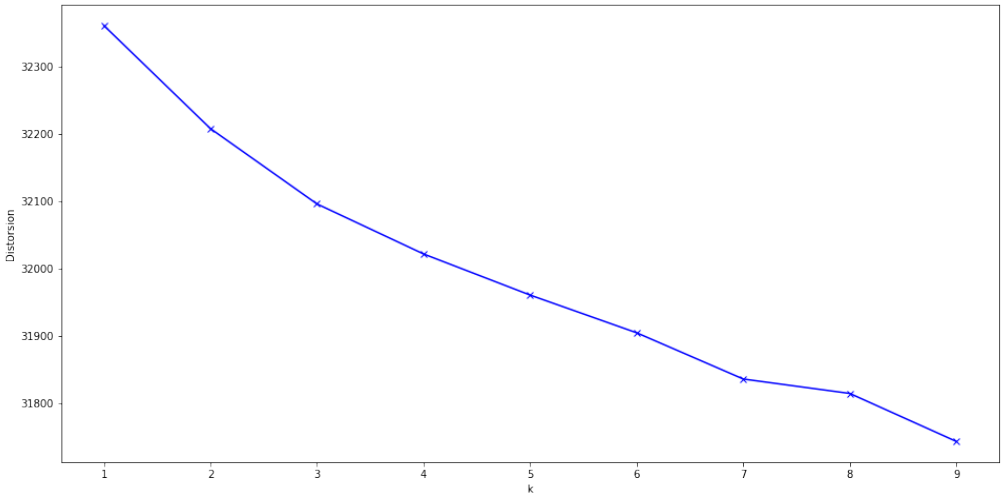

FIGURE 6.2 – Résultat de la méthode Elbow sur les *claims* des introductions

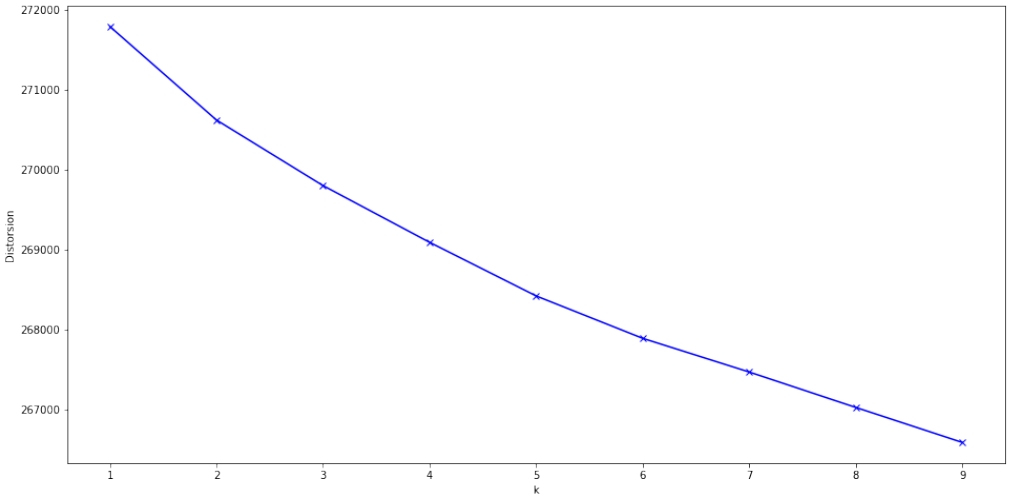

FIGURE 6.3 – Résultat de la méthode Elbow sur les *claims* des corps d'articles

6.4 Tableaux des scores des métriques d'évaluation

Note : Lem. pour Lemmatisation, CH pour Calinski-Harabasz, DB pour Davies-Bouldin, Eucl. pour Euclidienne, Manh. pour Manhattan

On remarque que les résultats en filtrant les mots outils sont légèrement meilleurs sur les corps d'articles. Toutefois, la différence est peu significative et ce filtrage n'est pas bénéfique sur les résumés, les introductions et les conclusions ; nous choisissons donc de garder les mots outils.

6.5 Graphiques liés à l'évolution diachronique des continents et au nombre d'autrices par continent

	Silhouette Score			CH score	DB score
	Cosinus	Eucl.	Manh.		
Lem., stopwords	0,005	**0,004**	**0,005**	**142,26**	8,83
Lem., pas de stopwords	0,007	0,003	0,002	127,7	8,71
Pas de lem., stopwords	0,005	0,002	0,002	119,48	13,89
Pas de lem., pas de stopwords	0,002	0,001	0,003	100,22	10,37

TABLEAU 6.1 – Résultats des métriques d'évaluation pour le clustering sur les résumés

	Silhouette Score			CH score	DB score
	Cosinus	Eucl.	Manh.		
Lem., stopwords	0,006	**0,008**	0,004	**144,7**	12,27
Lem., pas de stopwords	0,006	0,003	0,002	137,62	9,18
Pas de Lem., stopwords	0,007	0,003	0,008	121,16	12,7
Pas de Lem., pas de stopwords	0,005	0,002	0,025	113,48	13,07

TABLEAU 6.2 – Résultats des métriques d'évaluation pour le clustering sur les introductions

	Silhouette Score			CH score	DB score
	Cosinus	Eucl.	Manh.		
Lem., stopwords	-0,04	**0,007**	**0,003**	1047,74	11,14
Lem., pas de stopwords	-0,04	0,006	0,003	1125,18	9,23
Pas de Lem., stopwords	0,005	0,002	-0,027	827,14	14,93
Pas de Lem., pas de stopwords	0,005	0,002	-0,02	827,67	10,02

TABLEAU 6.3 – Résultats des métriques d'évaluation pour le clustering sur les corps d'articles

	Silhouette Score			CH score	DB score
	Cosinus	Eucl.	Manh.		
Lem., stopwords	0.003	0,001	-0,071	**1440,938**	11,617

TABLEAU 6.4 – Résultats des métriques d'évaluation pour le *clustering* sur toutes les parties mélangées

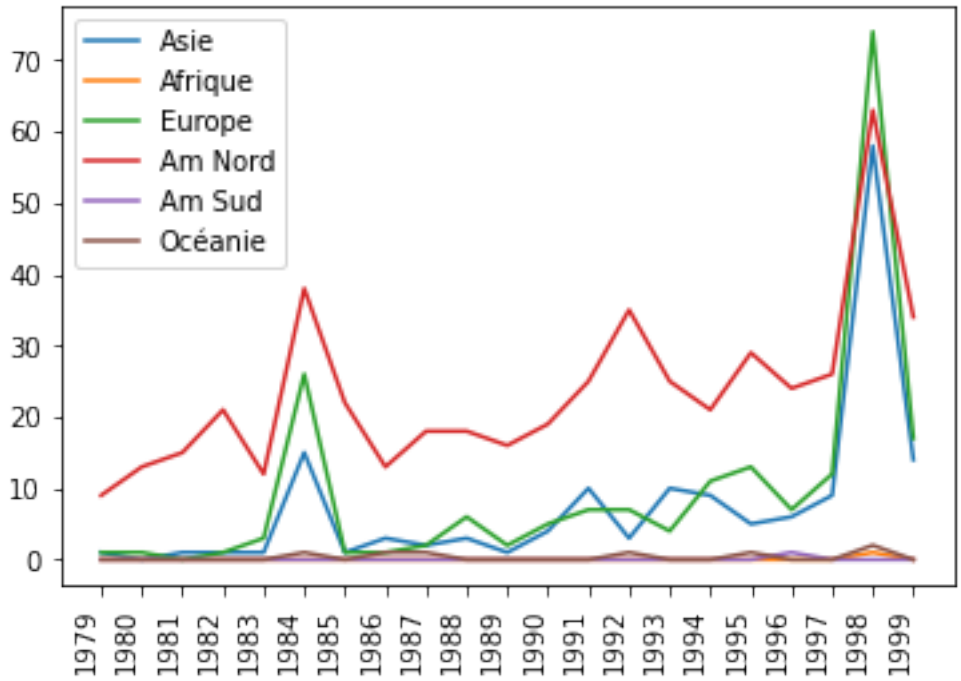

FIGURE 6.4 – Evolution diachronique (pré-2000) du nombre d'articles majoritairement affiliés aux différents continents

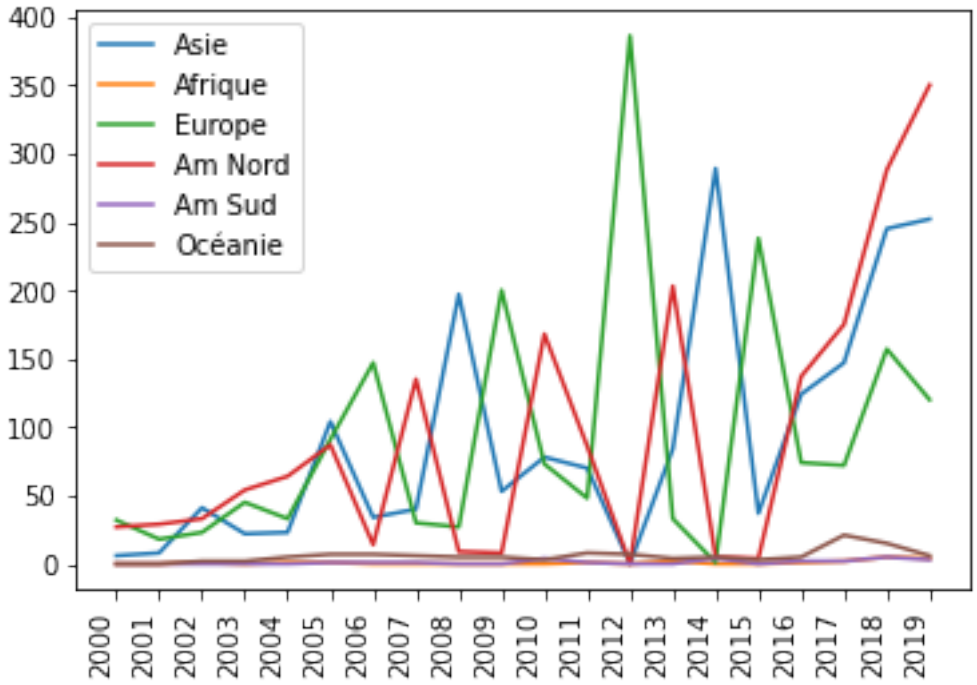

FIGURE 6.5 – Evolution diachronique (post-2000) du nombre d'articles majoritairement affiliés aux différents continents

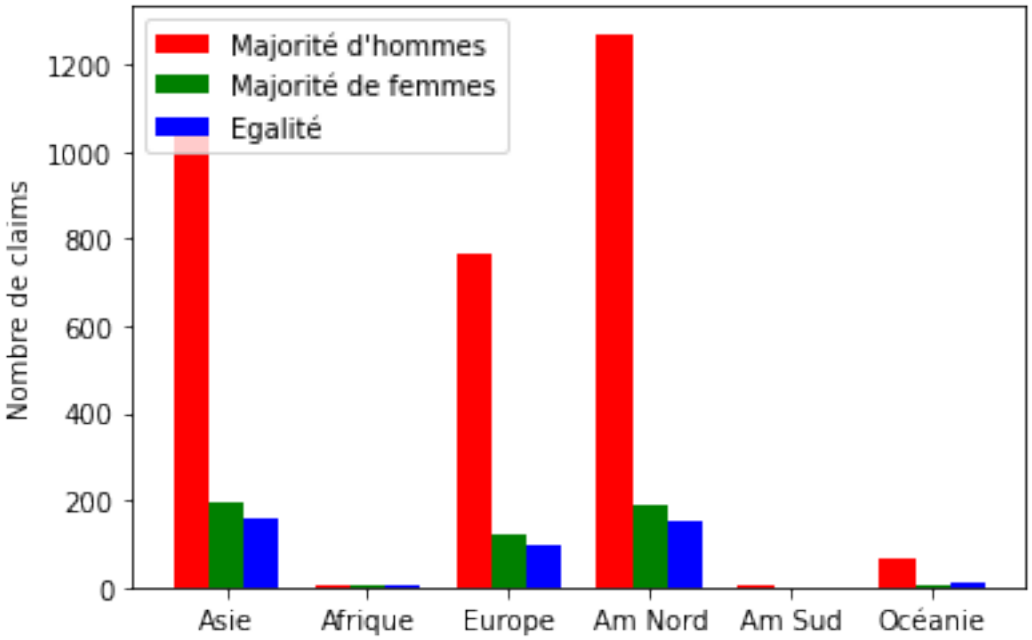

FIGURE 6.6 – Répartition des groupes d'auteur·ices selon les continents, nombres absolus

Bibliographie

Abdalla, M. and Abdalla, M. (2021). The grey hoodie project : Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Alt, C., Gabryszak, A., and Hennig, L. (2020). Probing linguistic features of sentence-level representations in neural relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1534–1545, En ligne. Association for Computational Linguistics.

Anonyme (2014). Computer ai passes turing test in 'world first'. *BBC*.

Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Mashuichi, H., and Ohe, K. (2009). TEXT2TABLE : Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192, Boulder, Colorado. Association for Computational Linguistics.

Augenstein, I. (2021). Determining the credibility of science communication. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 1–6, En ligne. Association for Computational Linguistics.

Bender, E. M. and Koller, A. (2020). Climbing towards NLU : On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, En ligne. Association for Computational Linguistics.

Blake, C. (2010). Beyond genes, proteins, and abstracts : Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2) :173–189.

Bossema, F., Burger, P., Bratton, L., Challenger, A., Adams, R., Sumner, P., Schat, J., Numans, M., and Smeets, I. (2019). Expert quotes and exaggeration in health news : a retrospective quantitative content analysis. *Wellcome Open Research*, 4 :56.

Bowman, S. (2022). The dangers of underclaiming : Reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 7484–7499, Dublin, Ireland. Association for Computational Linguistics.

Cabanac, G., Labbé, C., and Magazinov, A. (2021). Tortured phrases : A dubious writing style emerging in science. evidence of critical issues affecting established journals.

Caglayan, O., Madhyastha, P., and Specia, L. (2020). Curious case of language generation evaluation metrics : A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelone, Espagne (En ligne). International Committee on Computational Linguistics.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1) :1–27.

Case, C. M. (1927). *Scholarship in Sociology* in *Sociology and Social Research*. Number v. 12. University of Southern California.

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

Conway, M., Doan, S., and Collier, N. (2009). Using hedges to enhance a disease outbreak report text mining system. In *Proceedings of the BioNLP 2009 Workshop*, pages 142–143, Boulder, Colorado. Association for Computational Linguistics.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2) :224–227.

Ducel, F., Fort, K., Lejeune, G., and Lepage, Y. (2022). Do we Name the Languages we Study ? The #BenderRule in LREC and ACL articles. In *LREC 2022 - International Conference on Language Resources and Evaluation (LREC)*, Marseille, France.

Eissen, S. and Stein, B. (2002). Analysis of clustering algorithms for web-based search. In *International Conference on Practical Aspects of Knowledge Management*, volume 2569, pages 168–178, Vienne (Autriche).

Ethayarajh, K. and Jurafsky, D. (2020). Utility is in the eye of the user : A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, En ligne. Association for Computational Linguistics.

Facchinetti, R., Krug, M., and Palmer, F. (2012). *Modality in Contemporary English*. De Gruyter Mouton.

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

Fujii, A. and Ishikawa, T. (2001). Organizing encyclopedic knowledge based on the web and its application to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 196–203, Toulouse, France. Association for Computational Linguistics.

Ganter, V. and Strube, M. (2009). Finding hedges by chasing weasels : Hedge detection using Wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 173–176, Suntec, Singapour. Association for Computational Linguistics.

Ge, T., Zhang, X., Wei, F., and Zhou, M. (2019). Automatic grammatical error correction for sequence-to-sequence text generation : An empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6059–6064, Florence, Italie. Association for Computational Linguistics.

Graham, S. and Ghotra, T. (2021). Metric selection and promotional language in health artificial intelligence. In *medRxiv*.

Guillaume Cabanac, Cyril Labbé, A. M. (2022). "bosom peril" is not "breast cancer" : How weird computer-generated phrases help researchers find scientific publishing fraud. *The Bulletin*.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *ArXiv*, abs/2002.05651.

Holmes, J. (1982). Expressing doubt and certainty in english. *RELC Journal*, 13 :28 – 9.

Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Hyland, K. (1998). *Hedging in Scientific Research Articles*. Pragmatics & Beyond New Series. John Benjamins Publishing Company.

Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable NLP systems : How should we define and evaluate faithfulness ? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, En ligne. Association for Computational Linguistics.

Jin, Z., Chauhan, G., Tse, B., Sachan, M., and Mihalcea, R. (2021). How good is NLP ? a sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, pages 3099–3113, En ligne. Association for Computational Linguistics.

Khodak, M., Saunshi, N., Liang, Y., Ma, T., Stewart, B., and Arora, S. (2018). A la carte embedding : Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 12–22, Melbourne, Australie. Association for Computational Linguistics.

Kilicoglu, H. and Bergler, S. (2008). Recognizing speculative language in biomedical research articles : A linguistically motivated perspective. *BMC bioinformatics*, 9 Suppl 11 :S10.

Kilicoglu, H. and Bergler, S. (2010). A high-precision approach to detecting hedges and their scopes. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 70–77, Uppsala, Suède. Association for Computational Linguistics.

Koroleva, A. (2017). Vers détection automatique des affirmations inappropriées dans les articles scientifiques (towards automatic detection of inadequate claims in scientific articles). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, pages 135–148, Orléans, France. ATALA.

Koroleva, A. (2020). *Assisted authoring for avoiding inadequate claims in scientific reporting*. Theses, Université Paris-Saclay ; Universiteit van Amsterdam.

Koroleva, A., Kamath, S., Bossuyt, P., and Paroubek, P. (2020). DeSpin : a prototype system for detecting spin in biomedical publications. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 49–59, En ligne. Association for Computational Linguistics.

Koroleva, A. and Paroubek, P. (2017). On the contribution of specific entity detection and comparative construction to automatic spin detection in biomedical scientific publications. In *The Second Workshop on Processing Emotions, Decisions and Opinions (EDO 2017)*, Poznan, Pologne. Zenodo.

Koroleva, A. and Paroubek, P. (2018). Annotating spin in biomedical scientific publications : the case of random controlled trials (RCTs). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon. European Language Resources Association (ELRA).

Lakoff, G. (1972). *Linguistics and Natural Logic*, pages 545–665. Springer Netherlands, Dordrecht.

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

Lakoff, R. (1973b). Language and woman's place. *Language in Society*, 2(1) :45–79.

Lallich, S. and Lenca, P. (2015). Indices de qualité en clustering. In *Journée thématique : clustering et co-clustering*, Issy Les Moulineaux, France. Société française de classification.

Larson, B. (2017). Gender as a variable in natural-language processing : Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Li, B., Zhou, L., Wei, Z., Wong, K.-f., Xu, R., and Xia, Y. (2014). Web information mining and decision support platform for the modern service industry. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 97–102, Baltimore, Maryland. Association for Computational Linguistics.

Li, Y., Zhang, J., and Yu, B. (2017). An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop : Natural Language Processing meets Journalism*, pages 106–111, Copenhague, Danemark. Association for Computational Linguistics.

Light, M., Qiu, X. Y., and Srinivasan, P. (2004). The language of bioscience : Facts, speculations, and statements in between. In *HLT-NAACL 2004 Workshop : Linking Biological Literature, Ontologies and Databases*, pages 17–24, Boston, Massachusetts, Etats-Unis. Association for Computational Linguistics.

Luttenberger, L. and Vulinovic, K. (2018). Claim strength identification for detecting exaggerations in science news. *Text Analysis and Retrieval 2018 Course Project Reports*, page 75.

M. Salih, N. and Jacksi, K. (2020). State of the art document clustering algorithms based on semantic similarity. *Jurnal Informatika*, 14 :58–75.

Malouf, R. (2000). The order of prenominal adjectives in natural language generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 85–92, Hong Kong. Association for Computational Linguistics.

Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Mit Press. MIT Press.

Mariani, J. J., Francopoulo, G., and Paroubek, P. (2019). The NLP4NLP Corpus (I) : 50 Years of Publication, Collaboration and Citation in Speech and Language Processing. *Frontiers in Research Metrics and Analytics*, 3 :1–30.

Martinovic, M. and Strzalkowski, T. (1992). Comparing two grammar-based generation algorithms : A case study. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Newark, Delaware, USA. Association for Computational Linguistics.

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.

Mathur, N., Baldwin, T., and Cohn, T. (2020). Tangled up in BLEU : Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, En ligne. Association for Computational Linguistics.

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

Mercer, R. E., Di Marco, C., and Kroon, F. W. (2004). The frequency of hedging cues in citation contexts in scientific writing. In Tawfik, A. Y. and Goodwin, S. D., editors, *Advances in Artificial Intelligence*, pages 75–88, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Mohammad, S. M. (2020). Gender gap in natural language processing research : Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7860–7870, Online. Association for Computational Linguistics.

Nowlin, R., Wirtz, A., Wenger, D., Ottwell, R., Cook, C., Arthur, W., Sallee, B., Levin, J., Hartwell, M., Wright, D., Sealey, M., Zhu, L., and Vassar, M. (2021). Spin in abstracts of systematic reviews and meta-analyses of melanoma therapies : A cross-sectional analysis (preprint).

Park, S.-B., Tae, Y.-S., and Park, S.-Y. (2006). Self-organizing n-gram model for automatic word spacing. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 633–640, Sydney, Australie. Association for Computational Linguistics.

Paterson, R. and Vincent-Akpu, I. F. (2022). *Impostor Syndrome with Women in Science*, pages 83–98. Springer International Publishing, Cham.

Patro, J. and Baruah, S. (2021). A simple three-step approach for the automatic detection of exaggerated statements in health science news. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 3293–3305, En ligne. Association for Computational Linguistics.

Radev, D. R., Muthukrishnan, P., Qazvinian, V., and Abu-Jbara, A. (2013). The acl anthology network corpus. *Language Resources and Evaluation*, pages 1–26.

Rousseeuw, P. J. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 :53–65.

Rubin, V. (2006). *Identifying Certainty in Texts*. PhD thesis, Syracuse University.

Saputra, D. M., Saputra, D., and OSWARI, L. D. (2020). Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. In *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, pages 341–346, Sriwijaya (Indonésie). Atlantis Press.

Saurí, R. and Pustejovsky, J. (2012). Are you sure that this happened ? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2) :261–299.

Schaeffer, S. E. (2007). Survey : Graph clustering. *Comput. Sci. Rev.*, 1(1) :27–64.

Schmelzer, R. (2020). Machines that can understand human speech : The conversational pattern of ai. *Forbes*.

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., and Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, 2021(1).

Shindo, H., Munesada, Y., and Matsumoto, Y. (2018). Pdfanno : a web-based linguistic annotation tool for pdf documents. In *LREC*, Miyazaki (Japon).

Simonite, T. (2022). When it comes to health care, ai has a long way to go. *Wired*.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., and Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series : Materials Science and Engineering*, 336 :012017.

Szarvas, G. (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08 : HLT*, pages 281–289, Columbus, Ohio. Association for Computational Linguistics.

Szarvas, G., Vincze, V., Farkas, R., Móra, G., and Gurevych, I. (2012). Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2) :335–367.

Velldal, E., Øvrelid, L., Read, J., and Oepen, S. (2012). Speculation and negation : Rules, rankers, and the role of syntax. *Computational Linguistics*, 38(2) :369–410.

Vold, E. T. (2006). Epistemic modality markers in research articles : A cross-linguistic and cross-disciplinary study. *International Journal of Applied Linguistics*, 16 :61 – 87.

Xu, S., Li, H., Yuan, P., Wu, Y., He, X., and Zhou, B. (2020). Self-attention guided copy mechanism for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1355–1362, En ligne. Association for Computational Linguistics.

Young, V. (2011). *The Secret Thoughts of Successful Women : Why Capable People Suffer from the Impostor Syndrome and How to Thrive in Spite of It.* Crown.

Yu, B., Li, Y., and Wang, J. (2019). Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4664–4674, Hong Kong, Chine. Association for Computational Linguistics.

Yu, B., Wang, J., Guo, L., and Li, Y. (2020). Measuring correlation-to-causation exaggeration in press releases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4860–4872, Barcelone, Espagne (En ligne). International Committee on Computational Linguistics.

Zhang, Y. and Song, L. (2019). Language modeling with shared grammar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4442–4453, Florence, Italie. Association for Computational Linguistics.

Zhu, C., Qiu, X., Chen, X., and Huang, X. (2015). A re-ranking model for dependency parser with recursive convolutional neural network. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 1159–1168, Beijing, Chine. Association for Computational Linguistics.