Langues « par défaut » ? Analyse contrastive et diachronique des langues non citées dans les articles de TALN et d'ACL

Fanny Ducel[1], Karën Fort[2,3], Gaël Lejeune[1,4], Yves Lepage[5]
(1) STIH, Sorbonne Université, 75006 Paris, France
(2) LORIA, Université de Lorraine, 54506 Vandœuvre-lès-Nancy, France
(3) Sorbonne Université, 75006 Paris, France
(4) CERES, Sorbonne Université, 75006 Paris, France
(5) IPS, Université Waseda, 808-0135 Kitakyûsyû, Japon
ducelfanny@gmail.com, karen.fort@loria.fr,
gael.lejeune@sorbonne-universite.fr, yves.lepage@waseda.jp

RÉSUMÉ
Cet article étudie l'application de la #RègledeBender dans des articles de traitement automatique des langues (TAL), en prenant en compte une dimension contrastive, par l'examen des actes de deux conférences du domaine, TALN et ACL, et une dimension diachronique, en examinant ces conférences au fil du temps. Un échantillon d'articles a été annoté manuellement et deux classifieurs ont été développés afin d'annoter automatiquement les autres articles. Nous quantifions ainsi l'application de la #RègledeBender, et mettons en évidence un léger mieux en faveur de TALN sur cet aspect.

ABSTRACT
Contrastive and diachronic study of unmentioned (by default ?) languages in TALN and ACL

We study the application of the #BenderRule in natural language processing articles, taking into account a contrastive and a diachronic dimensions, by examining the proceedings of two NLP conferences, TALN and ACL, over time. A sample of articles was annotated manually and two classifiers were developed to automatically annotate the remaining articles. This allows us to quantify the extent to which the #BenderRule is applied and to show a slight advantage in favor of TALN.

MOTS-CLÉS : #RègledeBender, diversité linguistique, éthique.

KEYWORDS: #BenderRule, language diversity, ethics.

1 Introduction

Cet article s'inscrit dans la lignée des études appliquant les techniques de traitement automatique des langues (TAL) pour explorer le champ du TAL lui-même (Mariani *et al.*, 2019). Les buts peuvent être par exemple d'explorer les progrès du domaine, en identifiant les découvertes ou les sujets d'actualité (Mariani *et al.*, 2013, 2014; Buitelaar *et al.*, 2014), d'aborder des questions éthiques, par exemple le plagiat (Mariani *et al.*, 2016), ou encore d'interroger la fiabilité des méthodes utilisées, comme en traduction automatique (Marie *et al.*, 2021). Les implications sont donc à la fois éthiques et scientifiques. Nous nous intéressons aux langues étudiées dans les articles de TALN et d'ACL, et à leur nombre. La question des langues étudiées en TAL a été à nouveau soulevée récemment, mais elle a été posée assez tôt dans l'histoire de la grammaire ou de la linguistique. En TAL, même si la majorité

des recherches est encore restreinte à un nombre limité de langues, la situation semble évoluer. C'est ce que suggèrent Joshi *et al.* (2020) dans leur article sur l'inclusion et la diversité linguistiques dans la communauté. Se préoccuper de la diversité des langues étudiées implique au minimum qu'elles soient nommées. Cela peut paraître évident, mais il y a dix ans Bender (2011) faisait remarquer que nombre d'articles ne mentionnent tout simplement pas les langues étudiées. Elle encourageait donc les chercheurs à les nommer explicitement, même quand (et en particulier si) il s'agit de l'anglais. Huit ans après, dans un billet, elle rappelle à la communauté que l'« anglais n'est ni synonyme ni représentatif de toutes les langues » (Bender, 2019). Cette exhortation à mentionner les langues étudiées est devenue la #RègledeBender que l'on peut résumer ainsi : « Il faut toujours spécifier la ou les langues étudiées ». En effet, ne pas citer la ou les langues étudiées pose non seulement problème au niveau linguistique (Mel'čuk, 1988, p. 4), mais amplifie également les inégalités (Hovy & Spruit, 2016; Bender *et al.*, 2021). Il n'existe cependant pas à notre connaissance de travaux publiés abordant explicitement la question de l'application de la #RègledeBender. Les études abordant la diversité linguistique dans les conférences, telles que (Mariani *et al.*, 2014) ou (Joshi *et al.*, 2020), ne traitent pas du sujet des langues par défaut. L'article présent est donc une première tentative empirique de quantification de l'application de la #RègledeBender. Pour mener à bien cette étude, nous entraînons deux classifieurs pour déterminer si les articles respectent ou non la #RègledeBender. Nous notons par la même occasion la ou les langues étudiées quand cela est possible.

Les contributions de cet article sont les suivantes : (i) nous avons assemblé un corpus de près de 8 500 articles issus des conférences TALN et ACL ; (ii) nous avons annoté manuellement un échantillon de 550 de ces articles, c'est-à-dire que nous avons vérifié si la #RègledeBender est appliquée ou non, et pris note des langues étudiées lorsque cela était possible ; (iii) nous avons analysé les résultats dans une perspective contrastive et diachronique et extrait les tendances générales.

2 Méthodologie

Corpus Nous avons assemblé un corpus d'articles en français (à l'exclusion donc des articles en anglais) provenant de la conférence TALN de 2000 à 2020, et un corpus d'articles en anglais d'ACL, de 1979 à 2020. Pour TALN, nous avons utilisé les articles au format texte de Boudin (2013), et pour ACL ceux fournis par l'*AAN Anthology Network Corpus* (Radev *et al.*, 2013) et avons OCRisé les années suivantes. Ce corpus comprend 8 434 articles : 1 172 pour TALN et 7 262 pour ACL.

Annotation manuelle L'un des auteurs du présent article (annotateur 0) a annoté manuellement 550 articles, 130 articles de TALN et 420 d'ACL, pris au hasard avec un équilibre par édition de la conférence. Chaque article a été affecté à l'une des quatre catégories suivantes et les langues étudiées ont été notées quand elles ont pu être identifiées :

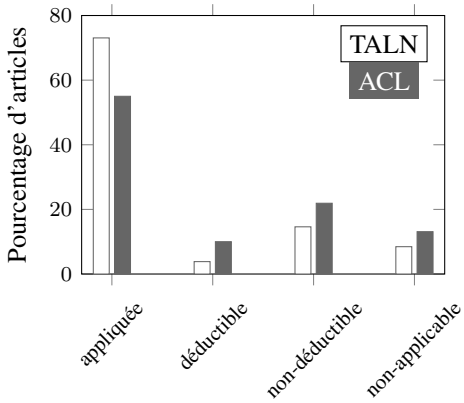
appliquée : la langue étudiée est clairement mentionnée dans l'article.

déductible : la #RègledeBender est applicable, non appliquée mais déductible des ressources citées. Nous utilisons la LREMap comme liste de ressources avec leur langue associée. [1]

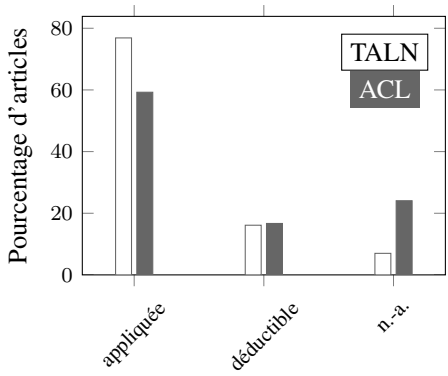
non-déductible : la #RègledeBender est applicable, non appliquée et non déductible. La langue n'est pas mentionnée et n'est pas facilement déductible d'un nom de ressource.

non-applicable : cas des articles qui ne traitent d'aucune langue en particulier.

1. Voir : `http://lremap.elra.info/`. Nous avons choisi cette référence afin de nous mettre à la place d'un lecteur lambda n'ayant pas accès à des bases de données plus importantes comme celle d'ELRA. A l'inverse, la LREMap, facilement accessible, est produite par un organisme faisant autorité dans le domaine des ressources langagières.



(a) Annotation manuelle (550 instances)



(b) Classification par $\text{Classif}_{\text{AA}}$ (8 434 articles)

FIGURE 1 – Annotation manuelle et classification automatique pour TALN (blanc) et ACL (noir).

Pour l'identification des quatre catégories ci-dessus (i) nous avons d'abord effectué des annotations d'essai afin d'identifier des problèmes et (ii) nous avons ensuite conçu un guide afin d'assurer la cohérence des annotations pour chacune des catégories. Ce guide comprend les trois règles suivantes :

i. un article doit être classé en « appliquée » s'il mentionne la langue étudiée, même si cette mention n'apparaît qu'une seule fois en fin d'article ;
ii. la LREMap est la seule référence pour la classe « déductible » ;
iii. l'utilisation des exemples donnés dans l'article pour déduire la langue étudiée est exclue car (a) les exemples peuvent figurer dans des langues difficilement identifiables par un annotateur humain ou un classifieur automatique, et (b) les exemples ne sont pas toujours fiables en ce qu'ils sont parfois des traductions dans la langue de l'article, à partir de la langue étudiée.

Afin de vérifier la qualité des annotations manuelles, les trois autres auteurs ont annoté en double un échantillon de ces données (62 articles dont 26 pour TALN et 36 pour ACL), de sorte que 15 % des annotations ont été dupliquées. Les annotations en double ont permis le calcul de l'accord inter-annotateurs avec l'annotateur 0 : 85 %, 83 % et 84 % respectivement pour les trois annotateurs. En fusionnant ces trois annotations, on obtient un Kappa de Cohen (1960) de 84 % avec l'annotateur 0, ce qui est généralement considéré comme bon (Artstein & Poesio, 2008).

La figure 1a donne la distribution des différentes classes sur l'échantillon annoté manuellement. Elle montre que les articles de TALN respectent davantage la #RègledeBender que les articles d'ACL : 73 % contre 55 %. En effet, dans cet échantillon, 27 % des articles de TALN et 45 % des articles d'ACL ne citent aucune langue. Pour ACL, 32 % des articles auraient dû appliquer la #RègledeBender, ce qui représente presque un tiers des articles d'ACL annotés, à comparer aux 19 % de TALN, ce qui reste néanmoins important. Afin de couvrir tous les articles disponibles pour les deux conférences, nous avons développé deux classifieurs automatiques.

Classification à base de *pattern-matching* Le guide d'annotation donné ci-dessus stipule que la #RègledeBender est respectée lorsqu'au moins un nom de langue est trouvé dans le texte. Nous avons donc conçu un classifieur, $\text{Classif}_{\text{PM}}$, qui se base sur la détection de noms de langue provenant

d'une liste de plus de 500 noms de langues en français et en anglais [2]. Si l'un de ces noms est trouvé, l'article « respecte » la #RègledeBender (classe « appliquée ») et inversement. Les noms de langue sont cherchés après pré-traitement classique : suppression de la ponctuation, mise en minuscules et *tokenisation.* Si au moins un des *tokens* est un nom de langue de la liste, l'article est classé comme respectant la #RègledeBender (classe « appliquée »). Un article est classé en « déductible » s'il contient au moins une ressource listée dans la LREMap (Calzolari *et al.*, 2012) [3]. Bien sûr, nous mémorisons les langues trouvées, ainsi que le nombre de leurs mentions. Le traitement ci-dessus ne distingue pas les cas où la langue citée n'est pas la langue étudiée, le but étant d'isoler les articles ne citant aucune langue. Ce cas peut-être dû au fait que les auteurs n'ont pas appliqué la #RègledeBender, ou que la #RègledeBender ne peut être appliquée, car le travail est théorique ou méta-linguistique.

La liste de toutes les ressources linguistiques monolingues présentes sur le site LREMap a été extraite en utilisant des méthodes de *Web-scraping* fournies dans la bibliothèque Python `BeautifulSoup` avec diverses améliorations [4]. Les *tokens* des articles sont comparés aux noms et aux chemins des ressources linguistiques. Les articles restants sont groupés dans une classe (« n.-a. » dans le tableau 1 et la figure 1b) : #RègledeBender non-appliquée ou non-applicable. Le nombre de classes dans le traitement automatique passe donc à trois. Notons également que le classifieur ne différencie pas les articles fournissant des données linguistiques sans citer la langue étudiée (ils devraient appliquer la #RègledeBender mais ne le font pas) des articles qui ne citent aucun nom de langue parce qu'ils sont à portée théorique ou méta-linguistique (la #RègledeBender ne s'applique pas pour eux).

Classification à base d'apprentissage automatique L'approche reposant sur le *pattern-matching* n'identifie pas les cas où une langue est mentionnée alors qu'elle n'est pas la langue étudiée. En effet, certains articles mentionnent dans la conclusion des langues que les auteurs souhaiteraient étudier dans de futurs travaux. On trouve aussi dans l'introduction ou la section sur l'état de l'art de nombre d'articles la mention de langues étudiées par d'autres auteurs. Pour remédier à ces problèmes, nous construisons d'abord un classifieur de phrases (pas un classifieur *d'articles*). Il utilise des techniques d'apprentissage automatique supervisé et permet de mieux définir la classe « appliquée » (l'étiquetage pour la catégorie « déductible » reste basé sur le principe de recherche de chaînes de caractères via la LREMap). La première version du classifieur traite l'anglais. Elle est entraînée sur 2 625 phrases extraites des articles d'ACL annotés manuellement et contenant au moins un nom de langue. La seconde version traite le français. Elle est entraînée sur 836 phrases d'articles de TALN faisant majoritairement partie des articles annotés manuellement contenant au moins un nom de langue. Pour l'entraînement de ces deux versions, nous étiquetons comme « appliquée » toutes les phrases provenant d'articles annotés manuellement et classés comme tels, tandis que les phrases provenant d'articles considérés comme n'appliquant pas la #RègledeBender sont étiquetées comme « non appliquées ou non applicable » (n.-a.).

Nous utilisons CountVectorizer de `scikit-learn`, avec ses paramètres par défaut et une répartition entraînement-test de 70-30, pour vectoriser le corpus, et LogisticRegression comme algorithme de classification supervisé. Cette combinaison s'est en effet avérée être la meilleure en comparaison d'autres combinaisons. L'utilisation de la pondération tf·idf ou de n-grammes plus longs que les unigrammes n'ont pas amélioré les résultats. La régression logistique a produit les meilleurs résultats parmi tous les classifieurs testés : arbres de décision, *Random Forest,* classifieurs bayésiens et SVM.

2. Voir : https://github.com/ISO639/2

3. Nous avons ajouté à la liste les ressources linguistiques suivantes, que nous considérons comme suffisamment utilisées (pour l'anglais et le français) : Le Monde, France info, France inter, NYT, PropBank, Washington Post, Wall Street Journal, Brown Corpus, Le Robert, Switchboard.

4. Cf. tutoriel de T. Ujhelyi à l'adresse : https://data36.com/scrape-multiple-web-pages-beautiful-so

	TALN				ACL			
	Préc.	Rappel	F-mes.	Instances	Préc.	Rappel	F-mes.	Instances
$\text{Classif}_{\text{PM}}$								
Appliquée	0,818	0,947	0,878	95	0,729	**0,978**	0,835	231
Déductible	0,250	0,600	0,353	5	**1,000**	0,595	**0,746**	42
N.-a.	0,500	0,211	0,296	19	0,741	0,429	0,543	147
Micro moy.	0,681	0,746	0,699	130	0,760	**0,748**	**0,724**	420
$\text{Classif}_{\text{AA}}$								
Appliquée	0,833	0,947	0,887	95	**0,856**	0,974	**0,911**	231
Déductible	0,357	1,000	0,526	5	**1,000**	0,286	0,444	42
N.-a.	0,500	0,211	0,296	19	**0,752**	**0,741**	**0,747**	147
Micro moy.	0,696	0,762	0,712	130	0,834	**0,824**	0,807	420

TABLE 1 – Performance des classifieurs sur les deux corpus (meilleurs résultats par classe en gras).

Nous construisons alors un classifieur d'articles, appelé $\text{Classif}_{\text{AA}}$, qui classe les articles dans la catégorie « appliquée » s'ils contiennent au moins une phrase classée comme telle par le classifieur de phrases de la langue de l'article, anglais ou français. Typiquement, une phrase comme « Le statut morphologique des affixes en chinois [...] » (Tseng *et al.*, 2020) sera classée comme « appliquée, alors que « Ces annotations peuvent délimiter des composantes de signes, [...] ou une traduction dans une autre langue comme l'anglais. » (Bragg *et al.*, 2019) ne sera pas classée comme « appliquée ». [5]

3 Résultats

Performance des classifieurs Nous comparons les résultats des classifieurs avec l'annotation manuelle dans le tableau 1. Nous calculons la précision, le rappel et la F-mesure des classifieurs en considérant les annotations manuelles comme référence. $\text{Classif}_{\text{AA}}$ donne de meilleurs résultats que $\text{Classif}_{\text{PM}}$ sur les articles des deux conférences en termes de précision. Les résultats sont peu convaincants pour la catégorie « déductible » ; on pourrait remettre en cause notre définition de cette catégorie et notre choix de la LREMap comme seule référence puisqu'elle est forcément incomplète.

Résultats de la classification et étude contrastive $\text{Classif}_{\text{AA}}$ classe 40 % des articles d'ACL de notre corpus comme n'appliquant pas la #RègledeBender. En comparaison, sur les 420 articles annotés manuellement, 45 % n'appliquaient pas la #RègledeBender. $\text{Classif}_{\text{AA}}$ classe 23 % d'articles de TALN comme n'appliquant pas la #RègledeBender, à comparer aux 27 % dans l'annotation manuelle. Les articles de TALN semblent donc appliquer la #RègledeBender plus souvent que ceux d'ACL. Une première explication concerne les périodes considérées : deux fois plus longue pour ACL que pour TALN. Or, il y a quarante ans, les travaux multilingues étaient sans doute plus rares. En conséquence, à ACL, il pouvait être évident que le travail effectué portait sur l'anglais, rendant superflue la mention de la langue. La deuxième explication serait que dans TALN il est plus naturel de préciser la langue étudiée puisque c'est moins souvent la langue de rédaction de l'article. À l'inverse, bien que la majorité des articles soit rédigée en français à TALN, mais parce que le français est globalement

5. Respectivement : « *The morphological status of affixes in Chinese* » et « *These annotations can demarcate components of signs, [...] or a translation into another language like English.* » dans les articles originaux.

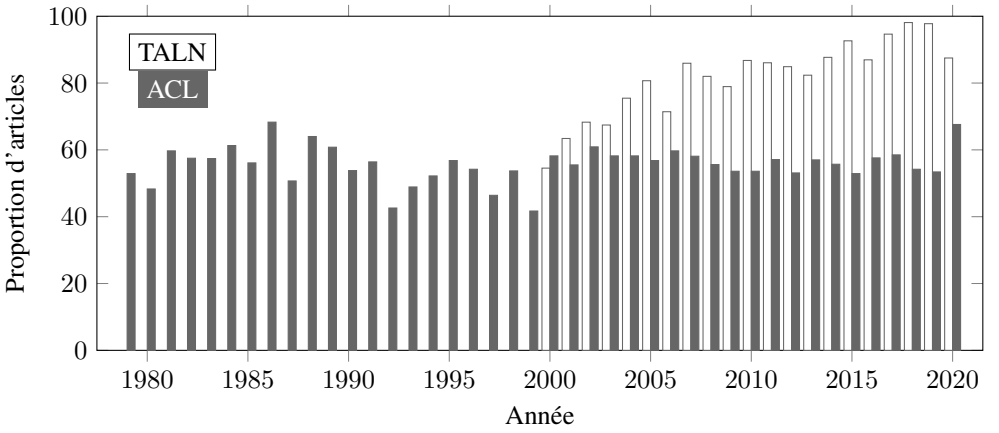


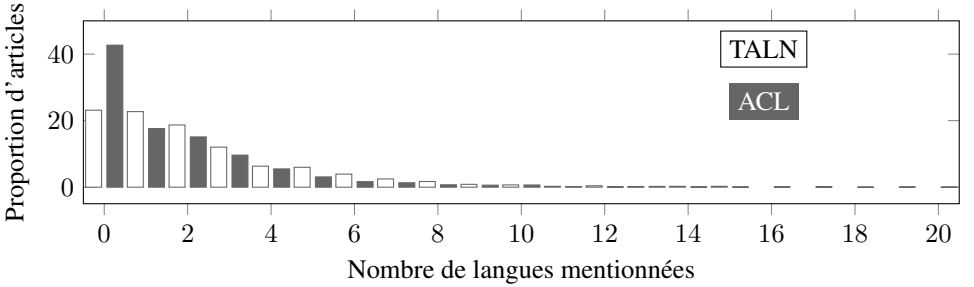
FIGURE 2 – Proportion d'articles appliquant la #RègledeBender par édition de TALN et d'ACL.

moins étudié que l'anglais, les auteurs ressentent peut-être le besoin de préciser qu'ils travaillent sur le français.

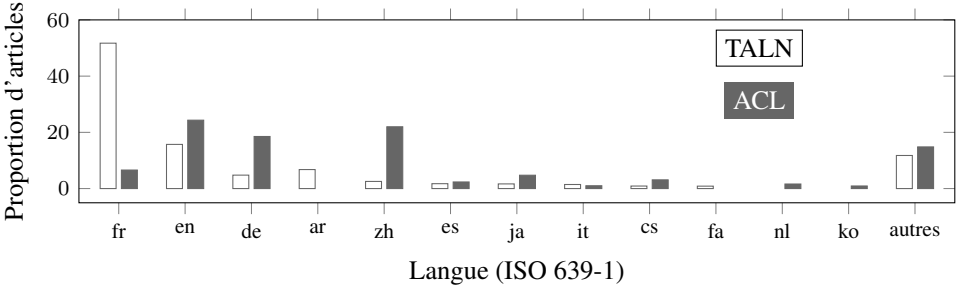
La figure 1b donne la distribution des articles par catégorie obtenue par classification automatique à l'aide de $\text{Classif}_{\text{AA}}$ pour l'ensemble des deux corpus TALN et ACL. Ces résultats reproduisent assez fidèlement la distribution obtenue par l'annotation manuelle (cf. figure 1a). La proportion de « déductible » double presque dans la prédiction automatique, ce qui réduit mécaniquement la proportion dans la dernière classe. L'inspection manuelle révèle qu'il y a de nombreux faux positifs ici. Parmi les articles appliquant la #RègledeBender, nous remarquons que certains ne citent qu'une seule langue et une seule fois. Pour ACL, cela concerne 13 % du total des articles, contre 4 % des articles annotés manuellement. À TALN on a seulement 4 % d'articles contenant une seule mention de langue (contre 3 % des articles annotés manuellement). On peut en conclure que les auteurs accordent plus d'attention aux langues qu'ils étudient dans la conférence TALN que dans ACL.

Étude diachronique La figure 2 montre la proportion d'articles appliquant la #RègledeBender pour chaque conférence et par année. Il ne semble pas y avoir de tendance chronologique marquante pour ACL, la proportion se situant presque toujours entre 50 % et 60 %. Néanmoins, nous observons une augmentation en 2020, après une tendance à la baisse en moyenne de 1986 à 2019. On peut se demander si cette augmentation n'a pas été provoquée par le billet de E. Bender en 2019. Pour TALN, la proportion d'articles appliquant la #RègledeBender oscille d'abord entre 54 et 86 % entre 2000 et 2013. Elle connaît ensuite une forte augmentation à partir de 2014, atteignant même les 98 % certaines années et ne descendant jamais sous les 86 %. On peut se demander si le premier billet d'E. Bender de 2011 n'a pas trouvé un écho important dans la communauté francophone.

Langues « par défaut » La figure 3a donne le nombre d'articles par nombre de langues mentionnées. Un grand nombre d'articles d'ACL ne citent aucune langue. Parmi les articles qui appliquent la #RègledeBender, on remarque qu'il y a une tendance plus forte à citer les langues lorsque l'étude n'est pas monolingue. En effet, dans les travaux multilingues, il est naturellement nécessaire de nommer les langues étudiées afin de les distinguer. Pour ACL, le système détecte beaucoup plus d'articles



(a) Proportion par nombre de langues mentionnées.



(b) Proportion des 10 langues les plus fréquemment citées dans chaque corpus (12 au total).

FIGURE 3 – Distribution des articles appliquant la #RègledeBender par nombre de langues mentionnées et par langues mentionnées.

qui n'appliquent pas la #RègledeBender. On constate également que les études monolingues y sont un peu plus représentées que les études bilingues, et que les études trilingues et quadrilingues sont encore moins nombreuses. Pour TALN, il y a presque autant d'articles qui ne citent aucune langue que d'articles qui en citent une. Cependant, nous remarquons que les articles bilingues sont très représentés, suivis de près par les articles trilingues.

La figure 3b montre que les 10 langues les plus représentées dans les deux conférences sont très similaires, différant par seulement deux langues : le coréen et le néerlandais n'apparaissent qu'à ACL, alors que l'arabe et le persan n'apparaissent qu'à TALN. Il convient de noter que presque toutes ces langues sont parlées dans des pays importants sur le plan politique ou économique, ou ont un nombre relativement important de locuteurs. À propos des langues les plus étudiées, E. Bender remarquait que « beaucoup des langues incluses sont parentes proches les unes des autres » (Bender, 2009). Ici, à l'exception du chinois, du japonais, de l'arabe et du coréen, elles appartiennent toutes à la famille des langues indo-européennes. En outre, deux ou trois langues seulement représentent plus de 10 % du total des langues mentionnées. Dans le corpus TALN, le français est sans surprise très majoritairement cité, alors que l'anglais constitue seulement 16 % des mentions. Pour ACL, le mandarin arrive juste après l'anglais. On peut imaginer que, dans quelques années, le mandarin dépasse l'anglais dans le classement. Cela ne signifiera pas forcément que cette langue sera plus étudiée que l'anglais, mais qu'elle sera davantage mentionnée. Les résultats sont bien sûr biaisés car la non-application de la #RègledeBender conduit à une sorte d'invisibilité du mot « anglais » (« français » pour TALN)

dans les articles. Si l'on extrait la liste des langues étudiées dans les articles qui n'appliquent pas la #RègledeBender (langues qui sont étudiées sans être mentionnées), on obtient uniquement l'anglais pour ACL et l'anglais et le français pour TALN. Cela confirme les affirmations d'E. Bender selon lesquelles l'anglais reste considéré comme une « langue par défaut », « synonyme de langue naturelle » et qu'il existe, en effet, une véritable « habitude de ne pas nommer la langue étudiée lorsqu'il s'agit de l'anglais » (Bender, 2019). Le cas du français pour TALN peut être interprété de façon similaire, ou en faisant l'hypothèse que les auteurs estiment évident qu'ils travaillent sur la langue dans laquelle ils écrivent.

4 Limites potentielles de l'étude

Il est bien entendu que nous n'avons pas lu l'intégralité du texte des articles que nous avons annotés manuellement, certaines annotations peuvent donc être incorrectes ou discutables. Cependant, l'accord inter-annotateurs obtenu est rassurant. Par ailleurs, la liste des langues que nous avons utilisée n'est pas exhaustive et nous avons probablement manqué des langues. Par exemple, l'article (Yu *et al.*, 2016) porte sur 107 langues que n'avons pas pu détectées, car il ne les liste que par leur code ISO, que nous ne prenons pas en compte. Un autre problème est le manque de précision concernant la langue, même lorsqu'elle est mentionnée. Par exemple, lorsque le chinois est mentionné, les auteurs ne précisent souvent pas de quel chinois il s'agit, même si, généralement, comme le dit E. Bender dans son billet, il s'agit du mandarin. Il en va de même pour presque toutes les langues.

Par ailleurs, nous avons dû exclure 24 articles en raison de problèmes d'OCR (fichiers illisibles). Comme nous n'avons pas corrigé manuellement les résultats de l'OCR, il est possible qu'il reste des problèmes non identifiés.

Enfin, nous avons limité notre étude aux articles jusqu'en 2020. Or, E. Bender a formulé sa règle en 2019. La proximité temporelle peut avoir une influence qu'il est difficile d'estimer aujourd'hui. Afin d'évaluer cet impact à plus long terme, les expériences rapportées ici sont à renouveler dans quelques années. Toutefois, notre objectif principal n'était pas d'évaluer l'impact du billet, mais la prévalence de la question.

5 Conclusion

Nous avons comparé l'application de la #RègledeBender dans deux conférences du TAL, TALN et ACL. Nous concluons que les articles de TALN mentionnent davantage les langues qu'ils étudient et qu'il y a une moindre représentation de la langue de la conférence dans les langues étudiés. Nous avons examiné l'évolution de l'application de la #RègledeBender au fil du temps mais n'avons pas observé de changement significatif pour ACL, même s'il y a un rebond en 2020. Enfin, dans les deux corpus, nous avons constaté que lorsque la #RègledeBender n'est pas appliquée, la langue étudiée est l'anglais ou le français pour quelques articles de TALN, ce qui confirme certaines affirmations d'E. Bender.

Le code et les ressources utilisés sont librement disponibles sur GitHub à fin de réplication. Pour évaluer le véritable impact du billet d'E. Bender, les classifieurs devront être rejoués dans les années à venir sur les articles de TALN et d'ACL (2021 et suivants). Une autre perspective est de reproduire l'expérience sur d'autres conférences internationales et nationales. Nous l'avons déjà fait sur les actes de LREC (Ducel *et al.*, 2022).

Références

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596. DOI : http ://dx.doi.org/10.1162/coli.07-034-R2.

BENDER E. (2019). The #BenderRule : On naming the languages we study and why it matters. *The Gradient*.

BENDER E. M. (2009). Linguistically naïve != language independent : Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics : Virtuous, Vicious or Vacuous ?*, p. 26–32, Athènes, Grèce : Association for Computational Linguistics.

BENDER E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, **6**(3).

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots : Can language models be too big ? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, p. 610–623, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3442188.3445922.

BOUDIN F. (2013). TALN archives : une archive numérique francophone des articles de recherche en traitement automatique de la langue. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles*, p. 507–514, Les Sables d'Olonne, France : Association pour le Traitement Automatique des Langues.

BRAGG D., KOLLER O., BELLARD M., BERKE L., BOUDREAULT P., BRAFFORT A., CASELLI N., HUENERFAUTH M., KACORRI H., VERHOEF T., VOGLER C. & RINGEL MORRIS M. (2019). Sign language recognition, generation, and translation : An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, p. 16–31, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3308561.3353774.

BUITELAAR P., BORDEA G. & COUGHLAN B. (2014). Hot topics and schisms in NLP : Community and trend analysis with Saffron on ACL and LREC proceedings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2083–2088, Reykjavik, Iceland : European Language Resources Association (ELRA).

CALZOLARI N., DEL GRATTA R., FRANCOPOULO G., MARIANI J., RUBINO F., RUSSO I. & SORIA C. (2012). The LRE Map. harmonising community descriptions of resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 1084–1089, Istanbul, Turquie : European Language Resources Association (ELRA).

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

DUCEL F., FORT K., LEJEUNE G. & LEPAGE Y. (2022). Do we name the languages we study ? the #BenderRule in LREC and ACL articles. In *Proceedings of International Conference on Language Resources and Evaluation (LREC) 2022*, Marseille, France : European Language Resources Association (ELRA).

HOVY D. & 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)*, p. 591–598, Berlin, Allemagne : Association for Computational Linguistics. DOI : 10.18653/v1/P16-2096.

JOSHI P., SANTY S., BUDHIRAJA A., BALI K. & CHOUDHURY M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6282–6293, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.560.

MARIANI J., FRANCOPOULO G. & PAROUBEK P. (2016). A study of reuse and plagiarism in speech and natural language processing papers. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, p. 72–83.

MARIANI J., PAROUBEK P., FRANCOPOULO G. & HAMON O. (2014). Rediscovering 15 years of discoveries in language resources and evaluation : The LREC anthology analysis. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande : European Language Resources Association (ELRA).

MARIANI J. J., FRANCOPOULO G. & 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. HAL : hal-02413751.

MARIANI J. J., PAROUBEK P., FRANCOPOULO G. & DELABORDE M. (2013). Rediscovering 25 years of discoveries in spoken language processing : a preliminary analysis of the ISCA archive. In *Annual Conference of the International Speech Communication Association*, Lyon, France. HAL : hal-01840831.

MARIE B., FUJITA A. & RUBINO R. (2021). Scientific credibility of machine translation research : A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 7297–7306, en ligne : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.566.

MEL'ČUK I. A. (1988). *Dependency Syntax : Theory and Practice*. New York : State University of New York Press.

RADEV D. R., MUTHUKRISHNAN P., QAZVINIAN V. & ABU-JBARA A. (2013). The ACL anthology network corpus. *Language Resources and Evaluation*, p. 1–26. DOI : 10.1007/s10579-012-9211-2.

TSENG Y.-H., HSIEH S.-K., CHEN P.-Y. & COURT S. (2020). Computational modeling of affixoid behavior in Chinese morphology. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 2879–2888, Barcelone, Espagne (en ligne) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.258.

YU Z., MAREČEK D., ŽABOKRTSKÝ Z. & ZEMAN D. (2016). If you even don't have a bit of Bible : Learning delexicalized POS taggers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 96–103, Portorož, Slovénie : European Language Resources Association (ELRA).