Représentation du genre dans des données open source de parole

Mahault Garnerin, Solange Rossato, Laurent Besacier

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Représentation du genre dans des données open source de parole

Mahault Garnerin[1, 2] Solange Rossato[2] Laurent Besacier[2]
(1) LIDILEM, Univ. Grenoble Alpes, FR-38000 Grenoble, France
(2) LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Avec l'essor de l'intelligence artificielle (IA) et l'utilisation croissante des architectures d'apprentissage profond, la question de l'éthique et de la transparence des systèmes d'IA est devenue une préoccupation centrale au sein de la communauté de recherche. Dans cet article, nous proposons une étude sur la représentation du genre dans les ressources de parole disponibles sur la plateforme *Open Speech and Language Resource*. Un tout premier résultat est la difficulté d'accès aux informations sur le genre des locuteurs. Ensuite, nous montrons que l'équilibre entre les catégories de genre dépend de diverses caractéristiques des corpus (discours élicité ou non, tâche adressée). En nous appuyant sur des travaux antérieurs, nous reprenons quelques principes concernant les métadonnées dans l'optique d'assurer une meilleure transparence des systèmes de parole construits à l'aide de ces corpus.

ABSTRACT

Gender representation in open source speech resources [1]

With the rise of artificial intelligence (AI) and the growing use of deep-learning architectures, the question of ethics and transparency in AI systems has become a central concern within the research community. We address transparency and fairness in spoken language systems by proposing a pilot study about gender representation in speech resources available through the *Open Speech and Language Resource* platform. We show that finding gender information in open source corpora is not straightforward and that gender balance depends on other corpus characteristics (elicited/non elicited speech, speech task targeted). The paper ends with recommendations about metadata and gender information for researchers in order to assure better transparency of the speech systems built using such corpora.

MOTS-CLÉS : traitement automatique de la parole, corpus, genre, locuteurs, données open source.

KEYWORDS: speech processing, corpora, gender, speakers, open source data.

1 Introduction

L'utilisation généralisée de l'apprentissage machine (*machine learning*) a fait des données un enjeu majeur de l'industrie et de la recherche. Les systèmes ont besoin d'une grande quantité de données étiquetées pour "apprendre" à modéliser correctement la tâche adressée. Les corpus de données, à l'instar de la puissance de calcul des machines, sont devenus de plus en plus grands, nous faisant entrer dans ce qu'on appelle aujourd'hui le *big data*. Mais outre la taille des corpus d'apprentissage, les chercheurs et chercheuses s'intéressent à une autre caractéristique de ces masses de données : leur

1. Publication originale en anglais, à LREC 2020.

Volume 1 : Journées d'Études sur la Parole, pages 244–252.
HAL : hal-02798544.

qualité. La notion de qualité des données peut trouver différentes définitions (Cai & Zhu, 2015), nous soutenons dans cet article que l'une des qualités principales de ces corpus est leur *transparence*.

Questionnant l'impact de tels outils sur nos sociétés, plusieurs études se sont intéressées aux biais existant dans ces systèmes : un cas bien connu dans le domaine du traitement automatique des langues (TAL) est l'exemple des plongements de mots (*word-embeddings*), avec les travaux de Bolukbasi *et al.* (2016) et de Caliskan *et al.* (2017) qui montrent le caractère socialement construit des données, encapsulant représentations et structures de pouvoir, incluant de fait les stéréotypes de genre. Des biais sexistes ont également été mis à jour pour des tâches de traduction automatique (Vanmassenhove *et al.*, 2018), ainsi que dans des systèmes de reconnaissance faciale (Buolamwini & Gebru, 2018). Dans une étude précédente, nous avons questionné l'impact du déséquilibre entre les catégories de genre dans les données d'entraînement sur les performances d'un système de reconnaissance automatique de la parole (RAP), montrant que la sous-représentation des femmes entraînait un biais de performance du système pour les locutrices (Garnerin *et al.*, 2019).

Dans la continuité de ce questionnement sur les liens entre représentations, données, et impacts sociétaux, nous étudions dans cet article la représentation du genre au sein d'une plateforme ouverte rassemblant des ressources langagières pour développer des outils de TAL. L'objectif de cette enquête est tout d'abord, d'observer les répartitions entre les catégories de genre au sein des corpus de parole. Cette répartition est envisagée en termes de nombre de locuteurs et de locutrices mais également en termes de temps de parole disponible pour chaque catégorie. Dans un second temps, nous proposons une réflexion sur les pratiques générales de mise à disposition de ressources, en nous appuyant sur quelques recommandations issues de travaux antérieurs.

2 OpenSLR

Open Speech Language Resources [2] (OpenSLR) est une plateforme créée par Daniel Povey, ayant pour objectif de centraliser des ressources langagières accessibles et téléchargeables gratuitement pour aider au développement de systèmes de parole. OpenSLR héberge actuellement 83 ressources. [3] Ces ressources sont constituées d'enregistrements audio transcrits, de logiciels, ainsi que de lexiques et de données textuelles nécessaires à la création de modèles de langue. D'autres plateformes proposent également ce type de ressources, la plupart du temps payantes. Nous nous concentrons donc sur les corpus de parole disponibles sur OpenSLR en raison de leur libre accès pour étudier à grande échelle la représentation du genre dans les corpus de parole.

Parmi les ressources disponibles sur la plateforme, nous avons analysé les 53 ressources de parole. Nous n'avons pris en compte ni les versions multiples d'une même ressource ni les sous-ensembles de ressources (e.g. LibriTTS, étant inclu dans LibriSpeech). Dans le cas de multi-versions, seule la dernière version a été conservée (e.g. TED LIUM). Plusieurs ressources contiennent de la parole dans différents dialectes ou langues et nous étudions chaque langue séparément. Nous avons ainsi un total de 66 corpus, dans 33 langues différentes avec 51 variantes dialectales/accentuelles. Les types de discours sont également variés (parole élicitée et lue, émissions radiophoniques, TEDTalks, enregistrements de réunions, appels téléphoniques, livres audio, etc.), ce qui n'est pas surprenant, étant donné le nombre d'acteurs ayant contribué sur la plateforme. Nous étudions cet échantillon

2. http://www.openslr.org.
3. Dernière consultation au 14 novembre 2019

Info. dispo.		#corpus
Non		24 (36.4%)
Oui	metadata	9 (13.6%)
	indexed	28 (42.4%)
	paper	5 (7.6%)
Total	-	**66**

TABLE 1 – Disponibilité des informations concernant le genre dans les corpus OpenSLR.

Info. dispo.	#corpus
Nombre d'individus	40
Nombre d'énoncés	32
Durée de parole	5
Nombre total de corpus	**42**

TABLE 2 – Type d'information disponible en fonction du genre dans les 42 corpus contenant des informations genrée.

pour aborder la question de la représentation du genre dans les corpus de parole. [4] OpenSLR ne fournissant pas de format défini et n'ayant pas d'exigences explicites concernant les structures de données, les ressources présentes sont également un bon reflet des pratiques des créateurs et créatrices de ressources concernant les méta-données.

3 Méthodologie

Afin d'étudier la représentation du genre dans les ressources disponible pour le traitement de la parole, commençons par définir ce que nous entendons par genre. Le genre est entendu ici comme variable binaire (homme/femme). Néanmoins, contrairement aux critiques faites par les sociologues J. Stacey et B. Thorne, qui dénoncent une cooptation du terme de genre (Stacey & Thorne, 1985), si le genre est envisagé ici comme une propriété des individus, il n'en reste pas moins l'expression d'un rapport social qui structure les relations et se retrouve dans les données de parole, de façon plus ou moins marquée en fonction des modalités de recueil des données langagières. Nous sommes également conscientes que les identités de genre sont plurielles et dépassent ces deux catégories, mais nous n'avons trouvé aucune mention de locuteurs ou locutrices non-binaires au sein des corpus étudiés dans notre étude.

Suite au travail de Doukhan & Carrive (2018), analysant la représentation du genre dans les flux télévisuels français, nous avons voulu explorer les corpus OpenSLR en réutilisant leurs notions de "taux de présence" (nombre d'individus) et de "taux d'expression" (temps de parole) pour rendre compte de la représentation du genre dans les données. Après téléchargement, ces informations ont donc été extraites manuellement des corpus.

3.1 Informations sur les locuteurs et locutrices et absence de méta-données

La première difficulté rencontrée est l'absence générale d'information (cf. Table 1). La prise en compte du genre dans la technologie étant un sujet de recherche relativement récent, les données démographiques sur le genre ne sont, la plupart du temps, pas mises à disposition par les créateurs et créatrices de ressources. Ainsi, en plus des caractéristiques générales du corpus mentionnées plus loin (voir 3.3), il nous semblait important de renseigner dans notre tableau final, si des informations sur le

4. Notre étude de cas ne prétend pas être exhaustive et il serait nécessaire d'inclure des ensembles de données fournies par des agences de ressources telles qu'ELRA ou LDC pour généraliser nos conclusions.

genre étaient fournies en premier lieu et le cas échéant de quelle manière. Les différentes modalités de cet attribut sont : *paper*, si un article a été explicitement cité dans la ressource, *metadata* si un fichier de métadonnées a été inclus, *indexed* si le genre a été explicitement indexé dans les données ou si les données ont été structurées en termes de genre.

3.2 Informations sur les durées et homogénéité des données

La deuxième difficulté concerne le fait que les informations sur le temps de parole ne sont pas standardisées, rendant impossible la comparaison de temps de parole entre individus ou entre catégories de genre (cf. Table 2). Lorsque des informations de durée sont fournies, la granularité utilisée varie selon les corpus. Certains auteurs indiquent les temps de parole en heures (e.g. (Panayotov *et al.*, 2015; Hernandez *et al.*, 2018)), d'autres le nombre d'énoncés ou de phrases (e.g. (Juan *et al.*, 2015; Google, 2019)), la définition de ces deux termes n'étant jamais explicite. Nous avons également constaté qu'il n'y avait pas de cohérence entre la durée de parole et le nombre d'énoncés, ce qui exclut la possibilité d'approximer l'une par l'autre.

3.3 Corpus

Le résultat final de notre analyse se traduit par un tableau [5] présentant toutes les caractéristiques des corpus. Les caractéristiques étudiées sont les suivantes : l'identifiant de la ressource (*id*) tel que défini sur OpenSLR ; la langue (*lang*) ; le dialecte ou l'accent s'il est spécifié (*dial*) ; le nombre total de locuteurs et locutrices ainsi que leur nombre dans chaque catégorie de genre (*#spk*, *#spk_m*, *#spk_f*) ; le nombre total d'énoncés ainsi que le nombre total d'énoncés par catégorie de genre (*#utt*, *#utt_m*, *#utt_f*) ; la durée totale, ou temps de parole, ainsi que la durée par catégorie de genre (*dur*, *dur_m*, *dur_f*) ; la taille de la ressource en gigaoctets (*sizeGB*) ainsi qu'un label qualitatif (*size*, prenant sa valeur entre "grand", "moyen", "petit") ; le taux d'échantillonnage (*sampling*) ; la tâche de discours ciblée pour la ressource (*task*) ; le caractère élicité ou non de la parole (*elicited*) [6] ; le statut de la langue (*lang_status*) : une langue est considérée comme ayant peu (low-resource) ou beaucoup (high-ressource) de ressources. Le statut de la langue est défini d'un point de vue technologique (c'est-à-dire : y a-t-il des ressources ou des systèmes de TAL disponibles pour cette langue ?) Il est fixé à la granularité de la langue (d'où le nom), quel que soit le dialecte ou l'accent (si renseigné) ; l'année de la publication (*year*) ; les auteurs et autrices de la ressource (*producer*).

4 Analyse

4.1 Disponibilité des informations sur le genre

Parmi nos 66 corpus, 36,4% ne fournissent aucune information sur le genre des locuteurs et locutrices. Plus de 20% des corpus ne fournissent aucune information sur les locuteurs et locutrices, quelle qu'elle soit. La Table 1 résume le nombre de corpus pour lesquels des informations de genre ont été

5. Le tableau final et le script utilisé pour l'analyse sont disponibles à l'adresse suivante https://github.com/mgarnerin/openslr_gender_survey

6. Nous définissons comme données de parole non-élicitées, des données qui auraient existé sans la création des ressources (par exemple : TedTalks, livres audio, etc.), les autres données de parole sont considérées comme élicitées

fournies et, le cas échéant, l'endroit où celles-ci ont été trouvées. La procédure de recherche était la suivante : nous avons d'abord examiné le fichier de métadonnées (si existant) et dans le cas contraire, nous avons cherché si le genre était indexé dans la structure des données. Si aucune information n'était trouvée, nous avons cherché s'il existait un article décrivant les données.

La Table 2 indique les types d'information renseignées dans le sous-ensemble des 42 corpus contenant des informations sur le genre des locuteurs et locutrices. La plupart du temps, seul le nombre d'individus dans chaque catégorie est indiqué ; cinq corpus fournissent également le temps de parole pour chaque catégorie. De ce fait, nous n'avons pas pu étudier le taux d'expression de chaque catégorie, comme dans le travail de Doukhan & Carrive (2018), mais nous avons analysé le nombre d'énoncés lorsque renseigné. Il convient toutefois de rappeler que la notion d'énoncé n'est jamais définie dans les ressources (le découpage est-il syntagmatique ? basé sur les groupes de souffles ou du aux limites techniques du système ?), il n'existe donc pas de cohérence entre nombre d'énoncés et temps de parole, et ces résultats sont à prendre avec prudence. En plus des 42 corpus pour lesquels nous avons réussi à trouver des informations sur le genre, nous avons recueilli manuellement ces informations pour 4 autres corpus, atteignant une taille d'échantillon finale de 46 corpus.

4.2 Genre et taux de présence

Parole élicitée vs non-élicitée. Lorsqu'on analyse le taux de présence de chaque catégorie de genre dans notre échantillon, la parité est atteinte avec 3 050 locutrices et 3 022 locuteurs. Cependant, certaines données sont pré-existantes à la création de ressources, notamment les données issues des médias, dans lesquels les femmes sont moins représentées (Macharia *et al.*, 2015). Le même résultat a d'ailleurs été mis en avant par l'étude de Doukhan & Carrive (2018). Nous avons donc croisé cette répartition avec le caractère élicité ou non de la parole, considérant comme non-élicitée toute parole qui aurait existé indépendamment de la création du corpus (e.g. TEDTalks, les interviews, les émissions de radio, etc.) Les résultats sont présentés dans la Table 3. Dans les deux cas (parole élicitée, respectivement non-élicitée), la différence entre les genres est relativement faible (5,6 points, respectivement 5,8 points), loin des 30 points de différence observés dans (Garnerin *et al.*, 2019). Une explication possible de cette observation est que les corpus, élicités ou non, restent le résultat d'un processus contrôlé, de sorte que la disparité homme/femme sera réduite autant que possible par les créateurs et créatrices des corpus. Cependant, on remarque qu'hormis Librispeech (Panayotov *et al.*, 2015), tous les corpus non élicités sont de petits corpus. En retirant Librispeech de l'analyse, nous observons un rapport femme/homme de 1/3-2/3, ce qui semble cohérent avec nos résultats précédents.

On peut donc conclure que la disparité de genre n'est observable que lorsque les données ne sont pas élicitées ou sciemment équilibrées. Cette représentation déséquilibrée n'est donc pas observée à l'échelle de l'ensemble de la plate-forme OpenSLR, la majorité des corpus étant élicités (89,1%). Ces résultats démontrent une volonté d'assurer la parité durant le processus de création des corpus.

"How can I help ?" : l'impact de la tâche. Lorsque les corpus de parole sont construits pour l'entraînement de systèmes ce sont la plupart du temps des systèmes reconnaissance de la parole (RAP) ou de synthèse vocale. En croisant la représentation du genre avec la tâche adressée, nous obtenons les résultats reportés dans la Table 4. Nous observons que si les taux de présence sont presque équilibrés au sein des corpus de RAP, les femmes sont mieux représentées dans les ensembles de données pour la synthèse. Cette observation fait écho au rapport de recommandation de l'ONU pour une éducation numérique égalitaire entre les sexes, qui indique qu'aujourd'hui la plupart des assistants vocaux ont une voix de femme en abordant les problèmes éducatifs et sociétaux que cela

Type de parole	#corpus	#F	#H
Élicitée	41	1782	1596
		52.8%	47.2%
Non-élicitée	5	1268	1426
		47.1%	52.9%
Non-elicitée	4	67	143
(sans Librispeech)		31.9%	68.1%

TABLE 3 – Taux de présence en fonction du type de parole

Tâche	#corpus	#F	#H
Reco.	12	2523	2615
		49.1%	50.9%
Synthèse	10	124	70
		63.9%	36.1%
NA	25	403	337
		54.5%	45.5%

TABLE 4 – Taux de présence en fonction de la tâche

	F	M
Nombre de loc.	591	551
	51.8%	48.2%
Nombre d'énoncés	72,280	143,342
	33.5%	66.5%

TABLE 5 – Nombre d'énoncés par catégorie de genre pour les 32 corpus fournissant ces informations. *N.B : deux corpus reportaient uniquement des nombres d'énoncés, le nombre de loc. est donc donné à titre indicatif*

soulève (West *et al.*, 2019). Cette conception genrée des assistants vocaux est parfois justifiée par des stéréotypes tels que "les voix féminines sont perçues comme plus serviables, plus sympathiques ou plus agréables". Les systèmes de synthèse vocale étant souvent utilisés pour créer des assistants vocaux, on peut supposer que l'utilisation de voix féminines est devenue pratique courante pour garantir l'adhésion du public au système. Cette affirmation peut toutefois être nuancée, notamment par les travaux de Nass & Brave (2005) qui ont montré que d'autres facteurs pouvaient justifier l'utilisation de voix féminines, tels que l'identification sociale et les stéréotypes culturels liés au genre.

4.3 Genre et taux d'expression

En raison d'un manque global d'informations sur le temps de parole, nous n'avons pas analysé le taux d'expression par catégorie. Cependant, le nombre d'énoncés est souvent renseigné, ou facilement retrouvable dans les corpus, et nous avons pu récupérer des fréquences par catégorie de genre pour 32 corpus. Si l'équilibre entre hommes/femmes est presque atteint d'un point de vue du taux de présence, les hommes sont plus représentés lorsqu'on s'intéresse au nombre d'énoncés (voir Table 5). Cependant, cette disparité n'est en réalité que l'effet de trois corpus contenant 51 463 et 26 567 (Korvas *et al.*, 2014) et 8376 (Hernandez-Mena, 2019) énoncés de locuteurs, alors que le nombre moyen d'énoncés par corpus est respectivement de 1942 pour les hommes et 1983 pour les femmes. Après avoir retiré ces trois valeurs extrêmes, la quantité de parole est équilibrée entre les catégories de genre. Le nombre élevé d'énoncés des trois valeurs extrêmes est cependant surprenant, ces trois corpus étant petits (2,1 Go, 2,8 Go) et moyens (5,2 Go). Cela met une fois de plus en évidence le problème de la notion d'énoncé (*sentence* ou *utterances*) qui n'est jamais explicitement définie. Une telle différence de granularité rend donc difficile la comparaison entre les corpus.

5 Recommendations

L'impact social du *big data* et les problèmes éthiques soulevés par les systèmes de TAL ont déjà été abordés dans des travaux antérieurs. Wilkinson *et al.* (2016) ont élaboré des principes pour la gestion des données scientifiques, les principes FAIR Data, basés sur quatre caractéristiques fondamentales des données qui sont la repérabilité (*foundability*), l'accessibilité (*accessibility*), l'interopérabilité (*interoperability*) et la réutilisabilité (*reusability*). Dans notre cas, la repérabilité et l'accessibilité sont prises en compte dès la conception, les ressources sur OpenSLR étant librement accessibles. L'interopérabilité et la réutilisabilité des données ne sont cependant pas encore atteintes. Une autre discussion sur la description des données au sein de la communauté du TAL a été initiée par Couillault *et al.* (2014), qui ont proposé une Charte sur l'éthique et les *big data* (*Ethics and Big Data Charter*), pour aider les créateurs et créatrices de ressources à décrire leurs données d'un point de vue juridique et éthique. Le travail de Hovy & Spruit (2016) a mis en évidence l'articulation complexe entre données, systèmes de TAL et leurs différentes implications sociales, avec, entre autres, les notions d'*exclusion*, de *surgénéralisation* et d'*exposition*. Plus récemment, les travaux de Bender & Friedman (2018) ont proposé la notion de *data statement* pour garantir la transparence des données. Nous espérons que la présente étude encouragera les chercheurs et chercheuses à décrire de manière exhaustive leurs ensembles de données, en suivant les lignes directrices proposées ci-dessus.

Sur l'importance des méta-données. La première conclusion de notre enquête est qu'il n'est pas facile d'obtenir une description exhaustive sur les locuteurs et locutrices dans les ressources de parole. Ce manque de méta-données est problématique d'un point de vue scientifique, car il empêche de garantir la généralisation des systèmes ou des résultats linguistiques basés sur ces corpus, comme le soulignent Bender & Friedman (2018), mais également éthique rendant impossible tout contrôle quant à l'existence d'une disparité de représentation pouvant conduire à des biais. Cette absence d'informations contextuelles sur la parole traduit aussi une conception du langage comme entité abstraite, plutôt que comme production située, qui mérite d'être questionnée (Hovy & Spruit, 2016).

Lorsque des informations sur la représentation du genre dans les données étaient fournies, celles-ci se portaient majoritairement sur le nombre de locuteurs et locutrices. Il serait intéressant d'avoir également accès à la durée des ensembles de données en heures ou minutes, globalement et par individu et/ou catégorie de genre. Taux de présence et taux d'expression n'étant pas égaux, l'un mesurant la représentation de chaque catégorie et l'autre la quantité de données disponibles. Des informations de durée standardisées pourraient permettre de vérifier rapidement l'équilibre entre les catégories de genre, sans s'appuyer sur une notion d'énoncé peu fiable. Lors de la collecte des données, nous avons remarqué que plus les ressources étaient récentes, plus il était facile de trouver des informations sur le genre, attestant de la visibilité croissante des thématiques de genre dans la technologie, mais si ce travail descriptif et important pour les futurs corpus, il doit également être effectué pour les ensembles de données déjà publiés, car ils sont susceptibles d'être utilisés à nouveau par la communauté.

Transparence dans l'évaluation. Le taux d'erreur-mots (WER pour *word error rate*) est généralement calculé comme la somme des erreurs commises sur l'ensemble des données de test divisée par le nombre total de mots dans la référence. Mais si une telle évaluation permet de comparer facilement les systèmes, elle ne tient pas compte de leurs variations de performance. Dans notre enquête, 13 des 66 corpus étaient accompagnés d'un article décrivant les ressources. Lorsque les performance des systèmes de RAP étaient reportées, aucune évaluation en terme de genre n'était faite, même si des informations sur la représentation du genre dans les données étaient renseignées. La communication

des résultats pour les différentes catégories est le moyen le plus simple de vérifier l'absence de biais dans les performances. Décrire ses données est un premier pas, mais pour une science ouverte et juste, l'étape suivante devrait être de prendre également en compte ces informations dans le processus d'évaluation. Un travail récent dans ce sens a été réalisé par (Mitchell *et al.*, 2019) qui a proposé de décrire les performances des modèles dans des "cartes modèles" (*model cards*), encourageant ainsi un rapport transparent des résultats.

6 Conclusion

Dans notre enquête sur le genre dans les corpus disponibles sur la plateforme OpenSLR, nous observons les tendances suivantes : la parité est globalement atteinte, mais les interactions avec d'autres caractéristiques des corpus révèlent que la disparité homme/femme nécessite plus qu'un simple nombre d'intervenants pour être identifiée. Dans les données non élicitées (c'est-à-dire toutes données de paroles qui auraient existé sans la création d'un corpus, comme les TEDTalks ou les émissions radiophoniques), nous avons constaté que, sauf dans le cas de Librispeech où l'équilibre entre les catégories de genre est contrôlé, les hommes sont plus représentés que les femmes. Il semble également que la plupart des corpus visant à développer les systèmes synthèse vocale contiennent principalement des voix féminines, peut-être en raison du stéréotype associant la voix féminine aux activités de *care*. Nous observons également que la description genrée des données a été prise en compte par la communauté, avec un nombre croissant de corpus fournis avec des métadonnées sur le genre au cours des deux dernières années. Notre échantillon ne contenant que 66 corpus, nous reconnaissons que nos résultats ne peuvent pas nécessairement être étendus à toutes les ressources linguistiques, mais cela nous permet relancer le débat sur les pratiques générales de description des corpus, soulignant le manque de méta-données, et d'actualiser le discours autour des implications sociales des systèmes de TAL.

Références

BENDER E. M. & FRIEDMAN B. (2018). Data statements for natural language processing : Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, **6**, 587–604.

BOLUKBASI T., CHANG K.-W., ZOU J. Y., SALIGRAMA V. & KALAI A. T. (2016). Man is to computer programmer as woman is to homemaker ? Debiasing word embeddings. In *Actes de NeurIPS 2016 (Neural Information Processing Systems)*, p. 4349–4357.

BUOLAMWINI J. & GEBRU T. (2018). Gender shades : Intersectional accuracy disparities in commercial gender classification. In *Actes de FAT 2018 (Fairness, Accountability and Transparency)*, p. 77–91, New-York City, USA : ACM.

CAI L. & ZHU Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science Journal*, **14**.

CALISKAN A., BRYSON J. J. & NARAYANAN A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, **356**(6334), 183–186.

COUILLAULT A., FORT K., ADDA G. & MAZANCOURT H. (2014). Evaluating corpora documentation with regards to the ethics and big data charter. In *Actes de LREC 2014 (Language Resources and Evaluation)*, p. 4225–4229, Reykjavik, Islande : ELRA.

DOUKHAN D. & CARRIVE J. (2018). Description automatique du taux d'expression des femmes dans les flux télévisuels français. In *Actes de JEP 2018 (Journées d'Études sur la Parole)*, p. 496–504, Aix-en-Provence, France.

GARNERIN M., ROSSATO S. & BESACIER L. (2019). Gender representation in French broadcast corpora and its impact on ASR performance. In *Actes de AI4TV 2019 (Workshop on AI for Smart TV Content Production, Access and Delivery)*, p. 3–9, Nice, France : ACM.

GOOGLE (2019). Crowdsourced high-quality UK and Ireland English Dialect speech data set. Web download at http ://www.openslr.org/83/.

HERNANDEZ F., NGUYEN V., GHANNAY S., TOMASHENKO N. & ESTÈVE Y. (2018). TED-LIUM 3 : Twice as much data and corpus repartition for experiments on speaker adaptation. In *Actes de SPECOM 2018 (Speech and Computer)*, p. 198–208, Leipzig, Allemagne : Springer.

HERNANDEZ-MENA C. D. (2019). TEDx spanish corpus. audio and transcripts in spanish taken from the tedx talks ; shared under the CC BY-NC-ND 4.0 license. Web Download.

HOVY D. & SPRUIT S. L. (2016). The social impact of Natural Language Processing. In *Actes de ACL 2016 (Volume 2 : Short Papers)*, p. 591–598, Berlin, Allemagne : Association for Computational Linguistics.

JUAN S. S., BESACIER L., LECOUTEUX B. & DYAB M. (2015). Using resources from a closely-related language to develop ASR for a very under-resourced language : a case study for Iban. In *Actes de INTERSPEECH 2015 (International Speech Communication Association)*, p. 1270–1274, Dresde, Allemange : ISCA.

KORVAS M., PLÁTEK O., DUŠEK O., ŽILKA L. & JURČÍČEK F. (2014). Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *Actes de LREC 2014 (Language Resources and Evaluation)*, p. 4423–4428, Reykjavik, Islande : ELRA.

MACHARIA S., NDANGAM L., SABOOR M., FRANKE E., PARR S. & OPOKU E. (2015). Who makes the news. Global Media Monitoring Project (GMMP).

MITCHELL M., WU S., ZALDIVAR A., BARNES P., VASSERMAN L., HUTCHINSON B., SPITZER E., RAJI I. D. & GEBRU T. (2019). Model cards for model reporting. In *Actes de FAT 2019 (Fairness, Accountability and Transparency)*, p. 220–229, Atlanta, GA, USA : ACM.

NASS C. & BRAVE S. (2005). *Wired for Speech : How Voice Activates and Advances the Human-computer Relationship*. MIT Press.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : an ASR corpus based on public domain audio books. In *Actes de ICASSP 2015 (Acoustics, Speech and Signal Processing)*, p. 5206–5210, Brisbane, Australie : IEEE.

STACEY J. & THORNE B. (1985). The missing feminist revolution in sociology. *Social problems*, **32**(4), 301–316.

VANMASSENHOVE E., HARMEIER C. & WAY A. (2018). Getting gender right in neural machin translation. In *Actes de EMNLP 2018 (Empirical Methods in Natural Language Processing)*, p. 3003–3008, Bruxelles, Belgique.

WEST M., KRAUT R. & EI CHEW H. (2019). I'd blush if I could : closing gender divides in digital skills through education.

WILKINSON M. D., DUMONTIER M., AALBERSBERG I. J., APPLETON G., AXTON M., BAAK A., BLOMBERG N., BOITEN J.-W., DA SILVA SANTOS L. B., BOURNE P. E. *et al.* (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, **3**.