Biais de genre dans un système de traduction automatique neuronale : une étude des mécanismes de transfert cross-langue

Guillaume Wisniewski* — **Lichao Zhu*** — **Nicolas Ballier**** — **François Yvon*****

* *LLF, Université Paris Cité & CNRS , 75 013, Paris France*
** *CLILLAC-ARP, Université Paris Cité, 75 013, Paris France*
*** *Université Paris-Saclay & CNRS, LISN, 91 403, Orsay France*

RÉSUMÉ. *Cet article a pour objectif de mettre en évidence les biais de genre dans les systèmes de traduction automatique et de rechercher leurs causes en étudiant les différentes manières dont l'information de genre peut circuler entre le décodeur et l'encodeur. Pour cela, nous décrivons un corpus minimal et contrôlé pour mesurer l'intensité de ces biais dans les traductions de l'anglais vers le français et du français vers l'anglais. Grâce à des méthodes de sondage et des interventions sur les représentations internes de l'encodeur, nos expériences montrent que l'information de genre est distribuée sur l'ensemble des représentations des tokens sources et cibles et que la sélection du genre en langue cible résulte d'une multiplicité d'interactions entre les diverses unités impliquées dans la traduction.*

ABSTRACT. *This paper describes a study on gender bias in French/English neural machine translation (MT) systems. We introduce a controlled corpus to measure the intensity of such biases in the two translation directions (from and into English). This corpus also allows us to investigate the information flow in a encoder-decoder architecture and to identify how gender information can be transfered between languages. Considering both probing as well as interventions on the internal representations of the MT system, we show that gender information is encoded in all token representations built by the encoder and the decoder and that there are multiple paths to transfer gender.*

MOTS-CLÉS : *biais de genre, traduction automatique neuronale, évaluation diagnostique en TAL.*

KEYWORDS: *Gender bias, Neural Machine Translation, Diagnostic Evaluation in NLP.*

TAL. Volume 63 – n° 1/2022, pages 37 à 61

1. Introduction

Il est largement admis (Callison-Burch *et al.*, 2006 ; Balvet, 2020) que les métriques automatiques telles que les scores BLEU ou METEOR sont inadaptées pour rendre compte des progrès de la qualité des traductions automatiques (TA) prédites par les systèmes neuronaux. Partant de ce constat, plusieurs protocoles ont été récemment proposés pour évaluer plus finement la traduction (Isabelle *et al.*, 2017 ; Burlot et Yvon, 2017 ; Burlot et Yvon, 2018). Ces protocoles reposent sur l'utilisation de jeux de tests élaborés (manuellement ou automatiquement) pour confronter les systèmes de TA à des problèmes de traduction spécifiques et bien caractérisés.

Les limitations de la traduction automatique ne se réduisent pas à leur incapacité à prendre en charge certains phénomènes linguistiques. Un autre problème important est l'existence de biais systématiques, en particulier de genre. Sous cette appellation, il faut distinguer plusieurs comportements problématiques : (a) le fait que des erreurs de traduction sont plus fréquentes pour des énoncés qui mettent en scène des actants de genre féminin ; (b) le fait que des traductions rendent linguistiquement explicite le genre des actants évoqués, alors que l'intention du locuteur peut être de le laisser ambigu ; (c) le fait que ces explicitations privilégient des assignations stéréotypiques, confortant, voire renforçant, des préjugés sexistes dans les textes traduits. Dans la typologie de Crawford (2017), affinée par Blodgett *et al.* (2020), ces problèmes sont susceptibles de fausser la manière dont certains groupes (ici, les femmes) sont représentés dans les textes *(representational harm)* ainsi que de conduire à un service (de TA) de moindre qualité pour les femmes *(allocational harm)*. Avec la massification de l'usage des technologies de TA, l'existence de tels biais est de plus en plus criante et dénoncée. Répondre à ces dénonciations exige à la fois des études précises (voir en particulier Savoldi *et al.* (2022) et les références citées), et des réponses appropriées de la part des fournisseurs de technologie[1].

Pour la paire de langues anglais français, ces problèmes peuvent être mis en évidence et quantifiés en observant la manière dont les marques de genre, qui peuvent être explicites ou non dans le texte source, se distribuent dans le texte cible. Une mesure de cet effet, mise en évidence dans plusieurs travaux, est proposée par Stanovsky *et al.* (2019), qui évalue les biais de genre à partir du décompte des erreurs portant sur la résolution d'anaphores pronominales. Ces travaux et d'autres études connexes visant à mesurer et à corriger les biais de genre sont passés en revue au § 2.

La première contribution nouvelle de cet article est d'étendre les analyses menées sur la traduction depuis l'anglais vers le français dans la direction inverse (§ 3). Nous proposons également de nouveaux contrastes pour mettre en évidence et quantifier les biais de genre grâce à la constitution d'un nouveau corpus contrôlé (§ 4). Nous nous intéressons, dans un second temps, à l'analyse des représentations internes d'un système de traduction neuronale à base de TRANSFORMER afin d'identifier plus finement la manière dont ces biais sont encodés dans les paramètres du réseau (§ 5). Nous pré-

1. Ainsi, les efforts récents de Google en la matière sont décrits dans ce billet.

sentons ensuite les premiers éléments d'une analyse causale permettant de mettre en évidence les différents mécanismes mis en œuvre dans le transfert de l'information de genre depuis le français vers l'anglais (§ 6).

2. Travaux connexes : mesurer et corriger les biais de genre

2.1. *Compter les erreurs et mesurer les biais*

La première étape pour étudier les biais de genre en TA consiste à les caractériser plus précisément, ainsi que les effets néfastes qu'ils peuvent produire auprès des utilisateurs de cette technologie (Blodgett *et al.*, 2020). Comme évoqué ci-dessus, ces auteurs distinguent en particulier les *biais de représentation*, qui conduiraient une TA à générer des textes véhiculant une représentation dénaturée des catégories sociales évoquées dans les textes traduits des *biais d'allocation*, qui se manifestent par un fonctionnement dégradé (des systèmes) pour certaines catégories d'usagers.

Lorsque l'on aborde ces questions sous l'angle quantitatif, à partir des observables que sont les sorties des systèmes de TA, deux situations sont à distinguer. Dans la première, le genre des personnes dont il est fait mention dans un texte source à traduire est indéterminé[2] et ne peut être déduit du contexte. Dans ce cas, on doit souhaiter que la traduction conserve cette ambiguïté, car tout autre choix impliquerait une interprétation non conforme aux intentions de l'auteur, tout en constatant que l'expression de cette ambiguïté est plus ou moins directe et transparente selon les langues, qui pour certaines disposent de formes neutres, ou bien ne marquent qu'exceptionnellement le genre, quand d'autres le marquent obligatoirement. À défaut, il semble souhaitable que les marques de genre qui seraient insérées le soient de manière équilibrée[3]. Lorsque ce n'est pas le cas, le système risque de créer, voire d'amplifier les biais de représentation, de fournir des informations faussées aux utilisateurs de la TA et de les propager dans les étapes de traitement ultérieures.

La seconde situation est celle dans laquelle l'information de genre[4] est explicite dans le texte source, auquel cas il est attendu qu'elle soit transférée correctement dans le texte cible, afin de toujours préserver les intérêts de l'auteur ainsi que celui des personnes qui seraient évoquées. Le système peut commettre deux types d'erreurs : (i) introduire dans le texte cible une ambiguïté qui est absente de la source ; (ii) se tromper dans l'expression du genre correct (complètement ou partiellement si le même genre est marqué sur plusieurs éléments du discours). En particulier entre dans cette

2. Cette formulation est simplificatrice, puisque, par exemple, il a longtemps été accepté en français dans certains usages que le genre masculin ait une valeur de générique – dans cette situation, il faudrait considérer que le genre des personnes représentées est indéterminé, alors même qu'une marque explicite de genre est présente.

3. Il est toutefois possible d'imaginer des situations ou des applications qui justifieraient de favoriser un genre (linguistique) plutôt qu'un autre dans les sorties.

4. Qu'elle soit encodée sous la forme d'une catégorisation binaire du genre ou bien qu'elle corresponde à des assignations plus fluides des identités de genre.

catégorie le fait de ne pas préserver l'ambiguïté ou la fluidité du genre alors que des pronoms sont disponibles pour éviter des assignations de genre binaire.

Même s'il est possible d'imaginer des situations dans lesquelles une traduction fidèle pourrait porter préjudice à certains usagers, il semble utile de mesurer les biais d'un système par des décomptes d'erreurs qu'il commet et la méthode que nous avons présentée *supra* s'inscrit dans cette démarche. Pour effectuer ces décomptes, la plupart des travaux analysant les biais de genre dans la traduction neuronale se sont concentrés sur le lexique de la profession (Kuczmarski et Johnson, 2018 ; Prates *et al.*, 2020), en étudiant aussi bien des corpus artificiels que des corpus réels (Gonen et Webster, 2020). Notons que la question du genre en TA peut être abordée sous d'autres angles : ainsi, Vanmassenhove *et al.* (2018) présentent des observations portant sur la distribution des verbes d'opinion en fonction du genre et du degré d'assertivité présumé chez les hommes et les femmes. Comme le montrent ces auteurs, qui étudient la traduction de dix langues vers le français, enrichir la phrase source (en anglais) par l'information explicite du genre du locuteur permet alors d'obtenir des traductions meilleures qu'un système qui ne dispose pas de cette information.

Une tentative de mesurer les biais dans la traduction vers l'anglais est détaillée par Cho *et al.* (2019), qui élaborent un indice du biais dans la traduction depuis le coréen *(translation gender bias index)*. Cet indice évalue la propension d'un système à traduire un pronom neutre en coréen en un masculin ou un féminin en anglais, ou bien encore en un groupe nominal non marqué pour le genre.

2.2. *Atténuer automatiquement les biais de genre*

Mesurer les biais permet aussi d'évaluer l'impact de travaux visant à les atténuer dans les TA. Ces travaux mobilisent principalement trois types de techniques (voir (Savoldi *et al.*, 2022) pour une étude récente). Une première consiste à manipuler les représentations lexicales. Elle est utilisée par Escudé Font et Costa-jussà (2019), qui injectent dans le système OpenNMT des plongements lexicaux entraînés avec l'algorithme *gender-neutral GloVe* (Zhao *et al.*, 2018). Ces auteurs testent ensuite la capacité à désambiguïser « *friend* » dans les traductions vers l'espagnol à partir des relations de coréférence ainsi que d'un nom de profession en attribut dans des phrases de la forme *I've known her for a long time, my friend works as a refrigeration mechanic.*

Les techniques de préannotation (Sennrich *et al.*, 2016) insèrent dans le texte source des marques explicites de genre, qui vont servir à orienter le système vers des traductions correctes. C'est, par exemple, l'approche suivie par Vanmassenhove *et al.* (2018), qui montrent que l'indication du genre des entités nommées dans l'anglais *(FEMALE Madam President, as a...)* permet d'améliorer les scores BLEU pour des traductions vers le français, l'italien, le danois et le finnois. Des résultats similaires sont obtenus par Basta *et al.* (2020) pour la traduction de l'anglais vers l'espagnol et des analyses complémentaires sont réalisées par Saunders *et al.* (2020). Cette tech-

nique est enfin utilisée par Kuczmarski et Johnson (2018) pour contrôler la traduction vers l'anglais de formes pronominales non marquées en turc dans des phrases telles que *« O bir doktor »* ou *« O bir hemşire »*.

Une troisième famille d'approche manipule les distributions des données d'apprentissage en s'appuyant sur des méthodes d'augmentation de données *(counterfactual data augmentation)*, CDA. Ainsi, Lu *et al.* (2020) engendrent automatiquement des corpus artificiels qui rétablissent l'équilibre en genre. Poursuivant cette direction, Saunders *et al.* (2020) montrent qu'il est plus simple et plus efficace de manipuler les distributions d'apprentissage en s'appuyant sur des méthodes d'adaptation au domaine. Ils utilisent un petit corpus artificiel équilibré en genre qui sert à adapter un système entraîné sur un corpus déséquilibré. Leur analyse de la traduction depuis l'anglais de trois langues montre que l'adaptation réduit les biais mesurés par les méthodes de Stanovsky *et al.* (2019).

3. Des jeux de tests contrôlés pour observer les biais de genre

Dans cette section, nous présentons la démarche suivie pour construire de nouveaux contrastes pour observer et quantifier les biais de genre en TA.

3.1. *Les corpus WinoGender et WinoBias*

Notre point de départ est l'étude de Stanovsky *et al.* (2019) qui formule des propositions concrètes pour évaluer les biais de genre, en s'appuyant principalement sur deux jeux de données : WinoGender (Rudinger *et al.*, 2018) et WinoBias (Zhao *et al.*, 2018), tous deux inspirés des schémas Winograd (Winograd, 1983)[5]. Un schéma Winograd repose sur une paire de phrases, chacune composée de deux propositions, qui ne diffèrent que d'un seul mot (ou une expression) prédicatif. Changer le verbe prédicatif induit un changement dans l'interprétation de la coréférence du pronom sujet dans la subordonnée, qui renvoie selon les cas soit au sujet soit à l'objet de la proposition principale, comme dans l'exemple suivant :

(1) *The city councilmen refused the demonstrators a permit because they [feared/advocated] violence*.

Dans cet exemple, la première partie de l'alternative (*feared*) conduit à interpréter *they* comme référant à *The city councilmen*, alors que la seconde (*advocated*) induit une coréférence avec *the demonstrators*. Ces schémas constituent des cas de test particulièrement difficiles pour les systèmes de TAL, car la résolution correcte de l'anaphore implique souvent une analyse profonde, voire des connaissances du monde.

5. On se reportera à Levesque *et al.* (2012) pour une discussion de ces schémas et à Amsili et Seminck (2017) pour leur adaptation au français.

Rudinger *et al.* (2018) décalquent ce schéma d'alternance pour cent vingt couples de phrases en mobilisant deux types de constructions pour constituer le corpus WinoGender :

(2) *[entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances].*

(3) *[entity1] [interacts with] [entity2] and then [interacts with] [pronoun] for [circumstances].*

Les jeux de tests qui en dérivent reposent alors sur l'établissement de la relation de coréférence entre *he* ou *she* et son antécédent dans des phrases comme (l'antécédent escompté est entre crochets) :

(4) *[The developer] built a website for the tailor because [she] is an expert in building websites.*

(5) *The developer built a website for [the tailor] because [he] wants to sell cloths online.*

WinoBias est construit sur des principes similaires et comprend un ensemble équilibré de 3 160 phrases contenant des anaphores pronominales dont l'antécédent est un nom d'activité ou de profession. L'association entre un pronom et un nom est également répartie entre (a) des situations « stéréotypiques » (conformes aux distributions par genre de ces activités dans la population) et non stéréotypiques ; (b) des structures dans lesquelles l'anaphore peut être résolue à partir de la syntaxe, et des structures pour lesquelles il faut des connaissances supplémentaires.

3.2. *Une évaluation des biais de genre*

Les 3 880 phrases issues de ces travaux sont utilisées par Stanovsky *et al.* (2019) pour mesurer les biais de systèmes traduisant depuis l'anglais vers huit langues dans lesquelles le genre est grammaticalisé.

L'exemple (4) ci-dessus correspond à une situation non stéréotypique et sera jugé correct si *developer* est traduit par *développeuse*, incorrect sinon. Selon ces auteurs, le biais se manifeste par des erreurs de traduction qui privilégient des genres associés à des rôles stéréotypiques, plutôt que ceux qui sont attendus au vu de la relation de coréférence. Ils proposent donc de les mesurer en comparant les taux d'erreur des traductions des pronoms associés respectivement à des noms masculins et féminins.

La méthode de Stanovsky *et al.* (2019) pour mesurer le biais est problématique à plusieurs titres, et finalement peu appropriée pour notre étude. En effet, outre son utilisation d'un corpus artificiel, elle repose sur un repérage automatique du genre du nom choisi par le système. Or ce repérage n'est pas toujours possible (quand le groupe nominal ne porte aucune marque de genre), comme lorsque l'on traduit la phrase suivante dans laquelle le genre du groupe nominal sujet ne peut être déterminé :

(6) *L'analyste se dirigea vers la caissière car elle devait faire un dépôt.*

Ce repérage peut également être bruité du fait d'erreurs dans l'étiquetage morphosyntaxique ou dans l'alignement automatique avec la phrase anglaise, ou encore du fait d'une traduction très erronée[6].

Sémantiquement, l'absence d'alternance de prédicat peut conduire à des ambiguïtés de la portée du pronom. À la différence des énoncés des tests de Winograd, dont la robustesse de l'interprétation fait l'objet de tests d'évaluation par des humains, les énoncés de WinoGender n'ont pas été testés et sont parfois potentiellement ambigus entre ce qu'on appelle parfois une interprétation *de dicto* ou *de re*. Ainsi, dans l'exemple ci-dessous, la proposition introduite par *because* peut expliciter soit un point de vue sur les propriétés du référent du sujet grammatical et constitue un commentaire du prédicat de la principale (interprétation *de dicto*) soit expliciter et constituer une propriété du référent de l'objet indirect en tant qu'actant du prédicat de la principale (interprétation *de re*) :

(7) *The chief gave the housekeeper a tip because she was satisfied.*

(8) *The chief gave the housekeeper a tip because he was helpful.*

Dans cet exemple, l'alternance prévue de la coréférence dans ce modèle de test n'est ainsi pas certaine : dans la phrase (8), *he* pourrait renvoyer à l'objet indirect *(de re)* ou au sujet grammatical *(de dicto)*, de sorte que l'alternance en genre de *chief* n'est pas garantie dans ce couple de phrases.

Un second problème est que ce test est difficile à « inverser » pour évaluer ces phénomènes dans la traduction du français vers l'anglais. Nos premières tentatives pour construire un jeu de tests en post-éditant des traductions automatiques de WinoGender se sont rapidement heurtées à de nombreux cas d'ambiguïtés dans la détermination du genre correct français. Il apparaît enfin que ce corpus contient un trop grand nombre de sources de variabilité (des structures de phrase et du lexique) pour que l'on puisse facilement manipuler et visualiser les représentations internes calculées pendant la traduction. Nous avons préféré utiliser dans nos expériences un jeu de données plus simple, en nous inspirant des travaux de Saunders et Byrne (2020) présentés à la section 2.2.

3.3. *Une évaluation plus contrôlée du biais de genre*

À l'instar de Saunders et Byrne (2020), nous avons construit, pour étudier le transfert de genre entre le français et l'anglais, un corpus parallèle sur les patrons (9-10) :

6. Ainsi, les trois résultats du tableau 1 qui portent sur les 3 880 exemples de WinoGender, excluent chacun plusieurs centaines de phrases (près de 900 pour le système `fairseq`), pour lesquelles le script d'analyse échoue à prédire le genre.

(9) `[DET] [N]` a terminé son travail. (p. ex. : `L'acteur` a terminé son travail.)

(10) The `[N]` has finished `[PRO]` work. (p. ex. : The `actor` has finished `his` work.)

Dans ces patrons, `N` est un nom de métier qui est soit masculin soit féminin (p. ex. en anglais $actor_M$/$actress_F$; en français, $acteur_M$/ $actrice_F$), `DET` est le déterminant français qui s'accorde avec le nom (soit sous la forme du féminin la_F, soit du masculin le_M, soit de l'épicène *l'*) et `PRO` est le pronom possessif anglais her_F ou his_M. Dans ces phrases, la seule marque de genre est alors portée, en anglais, par le pronom[7] possessif et en français par le groupe sujet.

Nous utilisons la liste complète des noms de métier français collectée par Dister et Moreau (2014) afin de saturer la position `[N]` en français et de sélectionner le déterminant correspondant. Cette liste contient les formes féminine et masculine de 1 696 professions. Pour 274 de ces noms, ces deux formes sont identiques (noms épicènes). Au final, notre corpus contient donc 3 392 phrases suivant le patron 9, et est parfaitement équilibré en termes de genre. Ces phrases ont été traduites automatiquement et vérifiées manuellement pour produire la liste des phrases correspondantes en anglais. L'ensemble des corpus ainsi construits est librement téléchargeable sur le site du projet Neuroviz[8].

La plupart des noms de métier identifiés par Dister et Moreau (2014) sont des noms composés rares qui sont absents dans le corpus d'entraînement de notre système d'apprentissage (cf. § 4.1) : comme le montre la figure 1(a), environ 30 % des noms de métier utilisés pour créer le corpus n'apparaissent pas dans le corpus d'entraînement et moins de 6 % d'entre eux apparaissent plus de 10 000 fois. L'observation, à la figure 1(b), de la distribution du nombre de tokens résultant de la décomposition en unités sous-lexicales[9] des noms de métier illustre également la faible fréquence à laquelle ceux-ci sont observés : seuls 574 noms de métier ont une fréquence suffisante pour donner lieu à un mot du vocabulaire du système de TA, tous les autres sont décomposés, le plus souvent en deux ou trois unités sous-lexicales.

4. Évaluation du transfert de genre entre le français et l'anglais

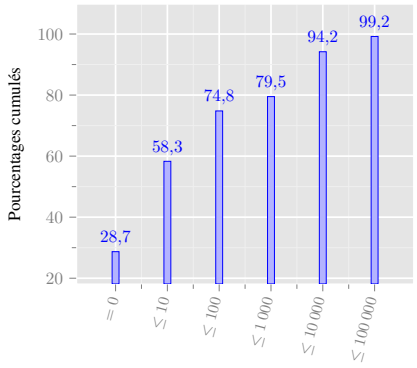
4.1. *Système de traduction*

Nous avons utilisé l'outil JOEYNMT, qui propose une implémentation « pédagogique » d'un système de traduction à base de TRANSFORMER (Vaswani *et al.*, 2017) permettant d'obtenir des résultats proches de l'état de l'art (Kreutzer *et al.*, 2019). Dans notre système, encodeur et décodeur sont composés de six couches, chacune

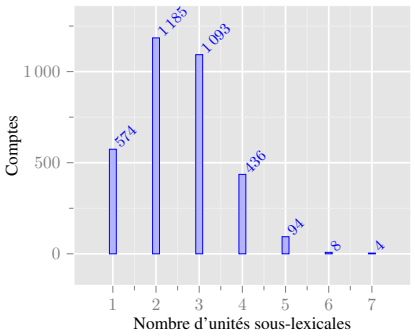
7. Nous suivons ici Huddleston *et al.* (2002) qui voient dans l'anglais une langue où le genre est peu grammaticalisé mais présent dans les relations de coréférence, comme avec les réfléchis *himself*, *herself* et *itself*.

8. `https://github.com/neuroviz/neuroviz/tree/main/gender_analysis_in_mt`.

9. Cette segmentation est détaillée à la section 4.1.



(a) Fréquences cumulées des noms de profession dans le corpus d'apprentissage.



(b) Distribution du nombre d'unités sous-lexicales dans les noms de métier.

Figure 1. *Rareté des noms de profession dans le corpus d'apprentissage*

avec huit têtes d'attention ; les couches de *feed-forward* comportent 2 048 paramètres et la dimension des plongements lexicaux est 512. Notre modèle comportait, au total, plus de 76 M de paramètres. Le système a été entraîné avec les données de la tâche « News » de la campagne WMT'15[10]. Les corpus `Europarl`, `NewsCommentary` et `CommonCrawl` sont utilisés pour l'apprentissage. Ils regroupent plus de 4 M de phrases et près de 141 M de tokens français. Tous les corpus ont été convertis en minuscules, tokenisés et segmentés en unités sous-lexicales en utilisant le modèle unigramme de l'outil SENTENCEPIECE (Kudo, 2018) et le vocabulaire résultant contient 32 000 unités. Le modèle est entraîné en optimisant l'entropie croisée à l'aide de la stratégie ADAM. Ce système obtient sur le corpus `newstest2014` un score BLEU de 34,0 pour la traduction du français vers l'anglais et de 32,7 pour la traduction de l'anglais vers le français.

Un autre point de comparaison est donné dans le tableau 1, qui reproduit pour ce système les mesures de biais de genre (Stanovsky *et al.*, 2019), en les comparant avec deux systèmes considérés dans cette étude, celui de `fairseq` (Ott *et al.*, 2018) et des traductions réalisées avec le système de Systran[11]. Il apparaît que notre implémentation de JOEYNMT atteint des performances conformes à celles des autres systèmes pour la prédiction du genre, avec une forte différence avec les prédictions pour le masculin et le féminin, et donc un fort biais de genre[12].

10. Il s'agit de la dernière campagne d'évaluation pour la traduction anglais français organisée dans le cadre de la conférence WMT (`http://statmt.org/wmt15`).

11. Dans ces deux derniers cas, nous utilisons les traductions de Stanovsky *et al.* (2019) et renvoyons à cette référence pour une description plus précise de ces deux systèmes.

12. *Précision* dénote la fréquence de la prédiction correcte du genre dans la phrase cible (complément du taux d'erreur à 1) ; ΔG_s dénote la différence de performances dans les traductions du masculin et du féminin.

	JOEYNMT	Fairseq	Systran
Précision	45,6	48,0	43,4
ΔG_s	30,1	4,4	41,8

Tableau 1. *Évaluation de la prédiction du genre de trois systèmes de traduction*

4.2. *Résultats expérimentaux*

Nous évaluons la capacité de notre système à prédire le genre des noms de métier en utilisant le jeu de tests décrit dans la section précédente et considérons comme point de comparaison les traductions engendrées par `e-translation`, le système de traduction développé par la Direction générale de la traduction de l'Union européenne[13].

4.2.1. *Mesure d'évaluation*

Nous nous intéressons, dans ce travail, à la capacité d'un système de traduction à transférer correctement l'information de genre du français vers l'anglais ou de l'anglais vers le français. C'est pourquoi notre évaluation ne porte que sur la capacité des systèmes à traduire correctement le groupe portant l'information de genre, indépendamment de la qualité de la traduction du reste de la phrase. Il est important de noter que, dans les deux directions, la position des mots portant l'information de genre est stable, ce qui simplifie l'évaluation de la correction des traductions par comparaison à celle de Stanovsky *et al.* (2019) qui implique une étape d'alignement automatique.

Vers le français, l'évaluation repose sur la prédiction du genre du groupe sujet traduisant `the [N]`, qui se trouve toujours en début de phrase. Quatre cas sont possibles, selon que le genre est porté par l'article et le nom (*la traductrice*), seulement par le nom (*l'actrice*), seulement par l'article (*la journaliste*), ou qu'il est complètement ambigu (*l'analyste*). Le tableau 2 décrit la distribution des différents cas dans notre corpus. Sauf mention contraire, nous considérerons que le genre du groupe sujet est correctement prédit lorsque le genre du déterminant et le nom sont tous deux corrects et évaluerons le taux de correction avec lequel le groupe sujet est prédit. Notons que cette mesure sous-évalue le nombre de phrases pour lesquelles le genre est correctement prédit : en effet, nous ne comptons comme « succès » que les phrases comportant la bonne traduction du nom de métier. Or, dans de nombreux cas, celui-ci est un mot rare (§ 3.3) et est mal traduit ; il est donc difficile d'évaluer si le genre est bien traduit ou non.

Dans l'autre direction, la vérification que le transfert du genre est correct en anglais est nettement plus simple et s'appuie sur le repérage du possessif (*her* ou *his*) dans la phrase cible. Il faut toutefois noter que pour les phrases sources dans lesquelles le déterminant et le nom de métier sont tous les deux épicènes (*l'analyste*) il est impossible de déterminer le genre et donc d'évaluer la qualité du transfert. Ce

13. `http://ec.europa.eu/cefdigital/eTranslation`

cas correspond à 272 exemples. Il est également envigeable qu'il ne soit pas possible de déterminer l'information de genre dans les traductions prédites par le système. Par exemple, dans certains cas, le système produit une traduction correcte n'utilisant pas les pronoms *her* ou *his* (*the programmer has finished working*) ; dans d'autres cas la traduction est complètement fausse (« L'inspectrice a fini son travail. » a été traduit en « *The young man bent on to work.* ») ou bien encore le possessif est traduit par *its* ou par *their*. Nous ne distinguons pas ces cas dans nos analyses et les regroupons tous sous la dénomination « autres » lorsque nous présentons les résultats.

4.2.2. *Résultats*

Pour la traduction du français vers l'anglais, il apparaît que notre système est capable de prédire correctement le pronom possessif dans seulement 52,4 % des phrases anglaises. *A contrario*, les résultats de `e-translation` atteignent presque la perfection : dans 90,9 % des hypothèses de traduction, le genre du pronom est correctement traduit. Pour la direction anglais-français, notre système est capable de prédire le déterminant correct pour 55,8 % des phrases et le nom pour 29,1 % des phrases. Ces observations suggèrent, qu'en plus de la difficulté de la tâche (traduction de noms de métier peu courants), le genre de la phrase est rarement correctement prédit. Quant à `e-translation`, il a été capable de prédire correctement le déterminant de 81,1 % des phrases et le nom de 57,8 % des phrases. La capacité de `e-translation` à correctement traduire les informations de genre a déjà été observée et celle-ci serait, *a priori*, due au choix de données d'apprentissage comportant majoritairement des énoncés sans biais de genre[14].

Pour la traduction vers l'anglais, le tableau 2 détaille ces scores : nous y indiquons, pour les différentes manières dont le genre peut être exprimé en français (§ 4.2.1), la proportion de phrases contenant *his* ou *her*. Ainsi, nous observons que, quand le déterminant et le nom ont tous les deux une forme spécifique au féminin, la traduction en *her* n'est choisie que dans 33,3 % des cas et *his* dans 18,5 % des cas. C'est d'ailleurs le seul cas, où le système génère plus souvent la forme féminine que la forme masculine. Ces résultats montrent que notre système, qui est entraîné à partir de corpus standard de traduction automatique, préfère nettement la traduction de *son* par un pronom masculin même dans des situations où il n'y a pas d'ambiguïté sur le genre du groupe nominal (p. ex. quand le déterminant et le nom ont tous les deux une forme spécifique au féminin). Dans l'ensemble, notre système n'obtient qu'une correction de 26,3 % pour les pronoms féminins, mais il prédit correctement le pronom dans 78,5 % des phrases au masculin. Ces conclusions corroborent celles obtenues par Saunders et Byrne (2020) sur les traductions de l'anglais vers l'allemand, l'espagnol ou l'hébreu. Des observations similaires ont été rapportées (Renduchintala et Williams, 2021) lorsque la traduction depuis l'anglais est testée sur un plus grand éventail de langues. À l'opposé, `e-translation` est capable de transférer correctement l'information de genre à partir du moment où celle-ci est marquée.

14. Le responsable du développement de `e-translation`, Markus Foti, a abordé cette question dans une interview accessible à ce lien.

Dét.	métier	nombre d'occurrences	pronom prédit	% phrases JoeyNMT	% phrases e-translation
l'	épicène	272	her	0,7 %	4,4 %
			his	80,1 %	94,1 %
			other	19,2 %	1,5 %
	fém.	251	**her**	**7,2 %**	**91,6 %**
			his	59,4 %	4,0 %
			other	33,4 %	4,4 %
	masc.	251	her	0,4 %	0,0 %
			his	**73,7 %**	**96,0 %**
			other	25,9 %	4,0 %
la	épicène	411	**her**	**31,6 %**	**93,0 %**
			his	43,9 %	0,3 %
			other	24,5 %	6,7 %
	fém.	898	**her**	**33,3 %**	**94,0 %**
			his	18,5 %	0,0 %
			other	48,2 %	6,0 %
le	épicène	411	her	0,7 %	0,0 %
			his	**84,4 %**	**95,4 %**
			other	14,9 %	4,6 %
	masc.	898	her	0,2 %	0,0 %
			his	**76,8 %**	**95,4 %**
			other	21,2 %	4,6 %

Tableau 2. *Pourcentage des hypothèses de traduction qui contiennent chaque type de pronom possessif selon la manière dont le genre est exprimé dans le sujet en français. Pour chaque cas de figure, le pronom anglais correct est en gras ; les scores de ces lignes correspondent alors à la correction du système.*

4.3. *Vers une analyse profonde du transfert du genre*

Les résultats de la section précédente mettent quantitativement en évidence les difficultés rencontrées par les systèmes de TA utilisés dans cette étude pour transférer l'information de genre entre langues, alors que celle-ci semble bien présente. Pour mieux comprendre les mécanismes à l'œuvre et mieux cerner les causes possibles de dysfonctionnements, nous poursuivons l'analyse en nous focalisant sur la direction français vers anglais, et donc sur les causes du choix d'un équivalent de traduction pour le mot français *son*. Ce choix est motivé par la structure systématique des phrases françaises et par la position relativement stable de ce mot dans la séquence d'entrée (il est toujours à la quatrième position à partir de la fin de la phrase), ce qui rend sa représentation interne assez facile à extraire et à manipuler. Notre principal objectif est alors de déterminer quels sont les éléments mis en jeu dans le choix du pronom anglais *his* ou *her* et d'étudier comment l'information de genre se propage dans le réseau pour construire les représentations numériques utilisées pour réaliser cette prédiction.

traduction	fréquence
_its	27,94 %
_his	**18,28 %**
_the	7,24 %
_her	**6,42 %**
_a	3,34 %
_their	2,92 %
_it	2,45 %
_sound	1,37 %
s	1,33 %
_he	0,76 %
autres	27,95 %

Tableau 3. *Les traductions les plus fréquentes du mot français* son *déduites des liens d'alignement de mots. Ainsi,* son *est aligné avec 3 658 types différents.*

Une première source de biais souvent mentionnée est liée aux données d'apprentissage. Comme reporté au tableau 3, il apparaît en effet que dans notre corpus d'apprentissage les traductions du pronom *son* par *his* sont trois fois plus fréquentes que les traductions par *her* (cette observation repose sur l'alignement des phrases sources et des phrases cibles avec `eflomal` (Östling et Tiedemann, 2016)). Nous pensons toutefois qu'il existe d'autres biais liés à la manière dont l'information de genre circule au sein de l'architecture TRANSFORMER.

Rappelons que, dans une architecture TRANSFORMER standard, le décodage s'effectue mot à mot de la gauche vers la droite. La sélection du mot anglais qui suit *has finished* s'appuie sur un vecteur contexte construit à partir des différentes couches d'attention de l'architecture. Trois mécanismes attentionnels sont simultanément à l'œuvre : l'auto-attention de l'encodeur, qui permet que les représentations des unités sources s'influencent mutuellement. L'auto-attention du décodeur, qui joue un rôle similaire côté cible, sous la contrainte que chaque unité n'ait accès qu'aux représentations des unités qui la précèdent. Enfin, l'attention croisée entre la source et la cible dans le décodeur, qui permet de contextualiser les représentations cibles en les combinant avec les représentations sources collectées sur la dernière couche de l'encodeur.

Pour traduire le genre du GN français, trois voies de propagation (non mutuellement exclusives) sont donc possibles : (a) une influence *directe* par le calcul de l'attention cross-langue ; (b) une influence *indirecte* passant par l'encodage (cross-langue) du genre dans la représentation du nom anglais, qui est propagée vers le pronom ; (c) une influence *indirecte* passant par l'encodage (monolingue) du genre dans la représentation d'autres mots français, en particulier du possessif français *son*, qui est ensuite propagée (cross-langue) vers le pronom anglais. Ces trois possibilités sont résumées dans la figure 2.

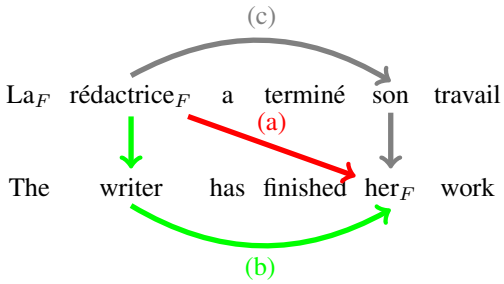


Figure 2. *Les différents mécanismes de transfert du genre du GN vers le pronom possessif en anglais*

Nous nous intéressons donc à démêler le rôle de ces différents mécanismes, par le truchement de plusieurs types d'analyses, qui vont permettre (a) de tester les représentations contextuelles en source et en cible par l'utilisation de sondes (section 5), une analyse que nous complétons en y intégrant des manipulations linguistiques (section 5.3) ; (b) de pondérer l'influence de ces trois mécanismes en utilisant des méthodes d'analyse causale (section 6).

5. Analyses par sondage

5.1. *Méthodes*

La première méthode considérée repose sur l'utilisation de sondes linguistiques *(probes)* (Belinkov et Glass, 2019) et consiste à tester la capacité à prédire le genre du GN français en observant seulement la représentation des mots sources construite par le système de TA. L'hypothèse sur laquelle repose cette méthode est que, s'il est possible de réaliser avec succès cette prédiction, c'est que les représentations correspondantes sont différentes pour les phrases comportant un GN masculin et pour les phrases qui comportent un GN féminin et peuvent donc influencer utilement le choix du pronom en anglais. Elles permettent de confirmer la possibilité d'une voie de transmission (b) et (c) de la figure 2.

En utilisant l'encodeur du système de traduction, nous calculons la représentation numérique associée à chaque mot du contexte source à droite du GN (soit : *a*, *terminé*, *son*, *travail*, . et `<eos>`), ainsi que le premier token (*the*) de la phrase en anglais dans toutes les couches du décodeur. Ces mots sont suffisamment fréquents pour donner lieu à une unique unité sous-lexicale et les représentations correspondantes sont à des positions stables. Nous considérons également une sonde qui est entraînée à partir de l'ensemble des tokens de la phrase cible (indépendamment de leur « valeur »), étant donné qu'il est impossible d'effectuer une analyse position par position en raison de la multitude des structures générées par la traduction automatique. Nous entraînons

ensuite un classifieur linéaire simple qui doit prédire le genre du GN à partir du vecteur représentant un token source ou un token cible.

Le classifieur utilisé est un modèle de régression logistique appris avec `scikit-learn` (Pedregosa *et al.*, 2011) en utilisant 75 % des données, et nous calculons le taux d'erreur sur les 25 % restant. Cette expérience est répétée cent fois pour pouvoir calculer l'intervalle de confiance de la mesure. Le classifieur est appris avec une régularisation ℓ_1 afin de contrôler la capacité des sondes (Hewitt et Liang, 2019).

5.2. *Résultats*

Les taux de correction *(accuracy)* obtenus par les diverses sondes sont dans le tableau 4. On observe en premier lieu que la représentation de *son* permet de prédire avec confiance le genre du nom, avec une correction supérieure à 80 %[15]. Comme la forme du mot n'est pas marquée en genre, on peut penser que cette information est apportée par le contexte qui va influencer la représentation interne. Cette influence du contexte se manifeste aussi par les corrections supérieures atteintes lorsque l'on exploite les couches les plus profondes de l'encodeur : la correction obtenue à partir des trois dernières couches de celui-ci dépasse les 90 %. Ce constat confirme qu'il existe bien un flux d'informations entre le pronom possessif et son antécédent, le nom de métier, ce qui correspond au chemin noté (c) dans la figure 2. On note que les scores de correction obtenus par la sonde lorsqu'elle prend en compte d'autres tokens sources sont également très élevés : ces scores sont comparables ou légèrement inférieurs à ceux obtenus avec *son*, ce qui montre que l'information de genre a un impact sur les représentations de tous les tokens sources, même lorsque ces tokens n'ont pas de relation syntaxique directe avec le groupe nominal sujet.

Les résultats de la sonde utilisant les représentations du décodeur comme caractéristiques (deux dernières colonnes du tableau 4) indiquent une tendance similaire : l'information de genre est encodée dans la représentation de *the*, même si le système génère toujours le même déterminant. Il apparaît également que la sonde est toujours capable de prédire le genre du nom de métier en français avec une assez grande correction à partir des représentations des tokens de la cible, illustrant, à nouveau, le caractère distribué de cette information.

Les résultats reportés dans la colonne *aléatoire* du tableau 4 correspondent à la prédiction d'une étiquette aléatoire à partir de ces mêmes représentations. Comme prévu, les résultats sont proches d'un tirage aléatoire et montrent que l'information présente dans les représentations est significative, puisque la sonde n'est pas capable d'apprendre des corrélations fallacieuses dans nos données (Hewitt et Liang, 2019).

15. Une analyse plus détaillée des prédictions montre que ce classifieur a un taux d'erreur similaire pour les phrases ayant un sujet féminin et pour les phrases ayant un sujet masculin.

	encodeur						aléatoire	décodeur	
couche	a	terminé	son	travail	.	eos	son	the	all
1	80,4	75,1	80,6	76,4	59,5	73,3	45,3	89,5	71,6
	±1,1	±0,3	±0,3	±0,6	±1,0	±1,0	±0,9	±0,2	±0,6
2	85,8	80,8	81,6	78,3	87,6	88,3	50,7	92,0	76,3
	±1,0	±0,2	±0,3	±0,7	±0,6	±0,7	±0,8	±0,1	±0,7
3	89,5	88,2	89,2	82,0	86,5	87,6	48,8	91,8	78,1
	±0,6	±0,2	±0,2	±1,1	±1,0	±0,6	±0,9	±0,1	±0,6
4	90,8	89,3	90,6	85,9	85,7	85,6	48,6	90,9	79,1
	±0,4	±0,2	±0,2	±0,9	±1,0	±0,7	±0,8	±0,2	±0,6
5	90,4	89,3	90,4	85,5	86,4	85,2	49,6	89,3	82,4
	±1,0	±0,2	±0,2	±0,8	±0,8	±1,2	±0,8	±0,2	±0,5
6	91,0	89,3	90,0	86,0	86,4	85,1	49,2	87,7	84,7
	±0,6	±0,2	±0,2	±1,0	±1,1	±0,8	±0,8	±0,2	±0,3

Tableau 4. *Correction (en %) de la prédiction du genre du syntagme sujet ou correction lorsque l'on cherche à prédire des étiquettes choisies aléatoirement à partir de la représentation de* son *(colonne aléatoire).*

5.3. *Sondage de représentations manipulées*

Nous étendons les tests par sondage de la section précédente en manipulant linguistiquement l'information de genre entre le déterminant, le nom, le pronom possessif et d'autres composantes de la phrase, afin de mesurer à quel point l'information de genre présente dans la sortie de l'encodeur est robuste à des variations du contexte d'occurrence. Le tableau 5 dresse la liste des manipulations considérées.

Les manipulations portent sur les caractéristiques suivantes. La première consiste à affaiblir le marquage du genre en remplaçant le DET (qui peut varier en genre) par *chaque*, qui est épicène. Nous varions ensuite le genre du nom déterminé par *son* : dans le patron initial, il est masculin (*travail*), nous proposons également une version avec un nom féminin (*activité*). Dans un troisième temps, nous augmentons la distance entre le groupe sujet et le possessif en insérant une proposition relative (*qui a chanté formidablement hier*) qui modifie le sujet. Cette manipulation est réitérée en insérant dans la relative un nom distracteur susceptible d'introduire du bruit dans la propagation du genre du groupe sujet (« homme » ou « femme »). Cet effet est encore renforcé lorsque l'on affaiblit le déterminant initial en le remplaçant par « chaque ». À l'inverse, nous considérons également la possibilité de renforcer le marquage du genre en introduisant un troisième composant adjectival dans le groupe sujet de manière à ce que le sujet en français soit toujours marqué explicitement.

La correction d'une sonde appliquée aux représentations des phrases manipulées est reportée dans le tableau 5. Une première observation est l'effet net des manipulations d'affaiblissement et de renforcement qui induisent respectivement des baisses et des hausses très sensibles de la qualité de la prédiction. Ces observations mettent en évidence le fait que l'encodage du genre dans les représentations en sortie de l'encodeur résulte d'un processus cumulatif dans lequel toutes les positions sources impliquées au sein du groupe sujet jouent un rôle effectif. À l'inverse, les autres manipula-

		encodeur					
	couche	a	terminé	son	travail	.	eos
Affaiblissement							
Chaque surveillant a terminé son travail.	1	73,1	73,6	65.7	63,5	53,9	56,7
	6	71,0	71,4	70,4	68,2	71,2	69,7
Renforcement							
Le surveillant *français* a terminé son travail.	1	99,9	98,5	95,0	80,6	62,0	80,4
	6	100,0	99,7	99,7	98,9	98,8	96,9
Genre du complément							
Le surveillant a terminé son *travail*.	1	79,4	74,6	79,0	75,0	58,8	72,0
	6	90,3	88,8	89,2	85,3	86,2	83,3
Le surveillant a terminé son *activité*.	1	80,5	75,5	78.6	62,6	57,6	67,2
	6	89,7	88,3	89,6	84,3	86,1	84,1
Éloignement							
Le surveillant *qui a chanté formidablement hier* a terminé son travail.	1	71,1	66,3	68,8	81,1	56,8	65,4
	6	91,5	91,0	90,5	86,8	81,2	82,1
Distracteur							
.sans affaiblissement							
Le surveillant *que cette femme critiquait* a terminé son travail.	1	65,7	66,6	69,3	79,50	62,8	68,5
	6	90,6	89,6	89,1	85,91	81,9	80,2
Le surveillant *que cet homme critiquait* a terminé son travail.	1	65,4	67,0	68,7	80,0	63,4	68,2
	6	90,3	89,3	89,7	86,6	81,0	79,9
.avec affaiblissement							
Chaque surveillant *que cet homme critiquait* a terminé son travail.	1	63,1	63,5	64,3	62,4	56,2	55,8
	6	72,1	71,4	69,7	69,9	71,8	69,2
Chaque surveillant *que cette femme critiquait* a terminé son travail.	1	63,3	64,6	65,9	63,4	55,4	55,2
	6	71,8	71,8	70,0	69,2	70,2	69,5

Tableau 5. *Correction des sondes pour les phrases transformées*

tions visant à dégrader l'encodage du genre, soit en insérant du matériel linguistique, soit en ajoutant des distracteurs, n'ont qu'un effet limité : si la qualité des sondes est globalement moins bonne lorsque l'on utilise la première couche, la correction des prédictions sur la dernière couche reste très stable, proche de 90 % pour les mots *a*, *terminé* et *son*. On voit ici à l'œuvre le mécanisme progressif par lequel les représentations contextuelles se construisent dans les différentes couches de l'encodeur : l'empilement de couches permet de filtrer les cooccurrences accidentelles au profit d'informations plus structurelles, moins dépendantes de l'éloignement entre mots.

En conclusion, cette première salve d'expériences a mis en évidence la présence d'une information distribuée portant sur le genre du groupe sujet en français, qui est disponible dans les sorties de l'encodeur et permet en théorie de prédire le genre du possessif en anglais avec une bonne correction (supérieure à 90 %). Or, les performances du système de traduction sont bien moindres, ce qui nous amène à explorer plus précisément, dans la section suivante, l'utilisation qui est faite de ces représentations dans le décodeur.

6. Éléments d'analyse causale

Dans cette section, nous nous attachons à mesurer si le genre est directement transféré depuis le groupe sujet ou bien si ce transfert passe aussi par les autres mots du contexte, dont nous avons vu (§ 5) qu'ils encodent cette information. Pour cette expérience, nous utilisons des techniques de manipulation des représentations internes du TRANSFORMER, en nous inspirant des méthodes d'analyse causale (Pearl, 2001 ; Vig *et al.*, 2020). L'analyse causale s'intéresse à quantifier les effets directs et indirects d'une variable X sur une variable Y, potentiellement médiatisés par une variable Z. Dans notre cas, on veut savoir si le genre grammatical du groupe sujet en français a un effet direct sur le choix de la forme du pronom possessif en anglais ou bien s'il y a un effet indirect du contexte de la phrase source ou du contexte cible.

6.1. *Une nouvelle mesure des biais de genre*

Nous commençons par décrire deux nouvelles mesures pour quantifier la préférence d'un système de TA pour une forme féminine ou masculine. Ces mesures reposent sur une comparaison de la probabilité d'engendrer la forme masculine ou féminine lors d'un décodage forcé[16] de la traduction de référence, plutôt que sur l'analyse de l'hypothèse de traduction prédite par le système de TA. Cette approche présente deux avantages : d'une part, elle n'est pas affectée par les éventuelles erreurs de la traduction automatique, d'autre part, comme elle s'appuie sur des comparaisons de probabilités d'engendrer certains tokens, elle permet de détecter et de quantifier des variations plus fines dans le modèle, même lorsqu'elles ne se traduisent pas par des modifications de l'hypothèse de traduction. Pour rendre plus concrètes nos explications et nos analyses, nous ne considérons ici que la traduction du français vers l'anglais, mais notre méthode se généralise à d'autres problèmes de traduction.

6.1.1. *Définitions*

Chaque exemple de notre corpus de test est représenté par un contexte de traduction source, $u(\texttt{DET},\texttt{N})$ où `N` et `DET` correspondent au nom du métier et à son déterminant. Le genre d'un exemple est noté $G(u)$ et peut prendre les valeurs masculin (M), féminin (F) ou indéterminé (I). Nos mesures reposent sur le calcul de $P(\text{his}|G(u))$ et de $P(\text{her}|G(u))$, soit la probabilité de produire *his* ou *her* sachant le contexte de traduction. Ces quantités sont impossibles à calculer exactement, puisqu'il faudrait sommer sur tous les débuts de phrases en anglais susceptibles de figurer devant le pronom. Nous réalisons donc une première approximation en considérant plutôt $P(\text{his}|G(u), e\triangleright)$, où $e\triangleright$ est le préfixe cible de la traduction de référence, et

16. Dans un décodage forcé, le i^{e} token prédit par le modèle n'est pas celui dont la probabilité est le plus élevée selon le modèle, mais celui correspondant au i^{e} token de la référence. Cette méthode de décodage permet de garantir que lors de la prédiction du pronom possessif anglais le contexte cible ne contienne pas d'erreur et de limiter ainsi le bruit dans nos mesures.

$P(\text{her}|G(u), e\triangleright)$ est défini de manière analogue. Ces deux quantités peuvent être calculées pendant le décodage forcé de la référence.

Nous considérons deux mesures du biais, dérivées respectivement de l'étude des contextes français pour lesquels le contexte permet de déduire le genre ($G(u) = M$ ou F), ou bien au contraire le laisse indéterminé ($G(u) = I$). Pour chacune des mesures, on s'intéresse aux valeurs moyennées sur tous les contextes u. La première mesure quantifie les biais de genre dans les contextes ambigus et est définie par :

$$b(u) = 1 - \frac{2 \times P(\text{his}|G(u) = I, e\triangleright)}{P(\text{his}|G(u) = I, e\triangleright) + P(\text{her}|G(u) = I, e\triangleright)} \quad [1]$$

Plus $b(u)$ est proche de 0, plus les probabilités de *his* et *her* sont proches, ce qui est attendu pour un contexte dans lequel le genre est indéterminé. Plus la probabilité d'engendrer *his* devient grande devant celle de *her*, plus $b(u)$ sera négatif. À l'inverse, une valeur proche de 1 indique une préférence pour la génération du pronom *her*.

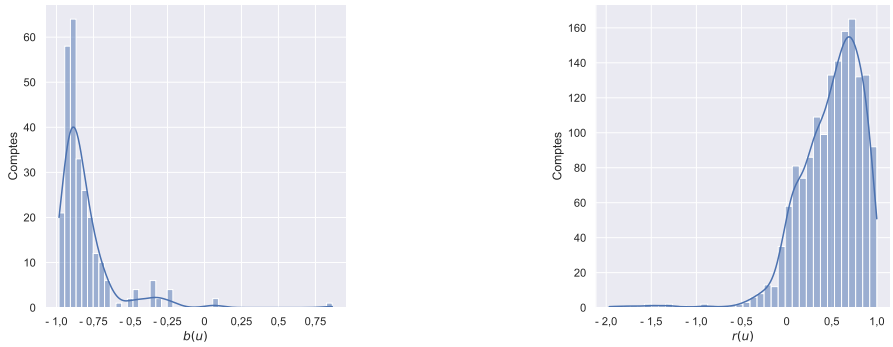
La seconde mesure est définie pour les contextes non ambigus par :

$$r(u) = 1 - \frac{P(\text{his}|G(u) = F, e\triangleright)}{P(\text{his}|G(u) = M, e\triangleright)} \quad [2]$$

Cette mesure compare, pour chaque contexte u, la probabilité d'engendrer le pronom *his* lorsque u correspond à la forme féminine du nom de métier, avec la probabilité d'engendrer *his* lorsque c'est la forme masculine qui est utilisée. La quantité $r(u)$ est proche de 0 quand les deux probabilités sont proches, c'est-à-dire lorsque la variation en genre du groupe sujet (le seul élément qui change dans le contexte) n'a pas d'influence sur les probabilités de sortie, ce qui manifeste une anomalie. Au contraire, si $r(u)$ est proche de 1, la probabilité de produire *his* lorsque le groupe sujet est masculin est très grande devant celle de le produire lorsque le groupe sujet est féminin, ce qui est le comportement souhaité ; une valeur fortement négative correspond à la situation (paradoxale) inverse : il est plus probable d'engendrer *his* avec un groupe sujet féminin plutôt qu'avec un groupe sujet masculin.

6.1.2. *Analyse des valeurs de base*

Nous avons représenté à la figure 3 la distribution des valeurs de b et de r calculées pour les phrases de notre corpus. Ces observations confirment les conclusions de la section 4.2 et montrent que notre système présente un biais très fort vers le masculin : pour les contextes indéfinis, la quasi-totalité des valeurs de $b(u)$ sont plus petites que $-\,0{,}75$ et cette mesure n'est positive que pour quelques rares contextes (p. ex. *autodidacte*, *aide* ou *interprète*), alors que la valeur moyenne de $b(u)$ devrait être nulle pour un système non biaisé. Pour les contextes non ambigus, $r(u)$ atteint une valeur moyenne de $0{,}46$, plus de 40 % des valeurs sont inférieures à $0{,}5$, et seules $0{,}7$ % des valeurs sont négatives. Ces observations montrent que la probabilité d'engendrer *his* est plus grande quand le groupe sujet est masculin que lorsqu'il est féminin, ce qui est attendu. En moyenne, passer d'un sujet masculin à un sujet féminin rend moins probable *his* d'un facteur 2 au bénéfice des alternatives : *her*, *it*, etc.



(a) Contextes ambigus (b) Contextes non ambigus

Figure 3. *Distributions de r et b sur les phrases de notre corpus*

6.2. *Importance du contexte dans la prédiction du genre*

6.2.1. *Trois manipulations des représentations*

Nous présentons maintenant une série d'expériences pour quantifier l'influence directe et indirecte des éléments du contexte source dans le choix de la forme du pronom possessif. Cette influence peut prendre diverses formes : en premier lieu par l'effet de l'attention cross-langue du décodeur qui peut intégrer directement des informations relatives au sujet français ; ce même mécanisme peut également s'appuyer indirectement sur les représentations contextuelles des mots d'autres sources, qui dépendent (*via* l'encodeur) du sujet ; la même dépendance indirecte peut être médiatisée par les représentations du début de phrase cible. Il peut enfin exister une influence directe entre le GN *cible* et le pronom qui s'exerce indépendamment du contenu de la source.

Pour essayer de démêler ces effets, nous construisons trois contextes alternatifs visant à neutraliser l'une ou l'autre des voies de transfert de l'information de genre identifiées dans la figure 2. Dans le premier contexte, noté u_1, la représentation lexicale de *son* est remplacée dès la première couche de l'encodeur par un plongement lexical « indifférencié » (dans nos expériences, nous avons choisi celui associé aux mots inconnus). Ce contexte réduit l'influence lexicale de *son*, sans empêcher toutefois que le genre soit transféré par la voie (c), puisque les couches supérieures de l'encodeur construisent une représentation contextualisée intégrant des informations sur le genre du groupe sujet. Les deux mesures $b(u_1)$ et $r(u_1)$ évaluent alors l'effet du transfert par les autres voies. Remarquons que même quand le token *son* est masqué dans la représentation de l'encodeur, il reste possible de calculer b et r, puisque ces deux mesures reposent sur un décodage forcé (qui produit toujours soit *his*, soit *her*).

Dans le second contexte (u_2), nous cherchons inversement à neutraliser l'effet du contexte source sur la représentation de *son* construite par l'encodeur. Pour cela, pour

chaque paire de phrases (forme féminine, forme masculine), nous substituons, dans chacune des phrases, la représentation de *son* sur la couche de sortie de l'encodeur par la moyenne des représentations construites pour ce token dans des contextes féminin et masculin. Nous supprimons ainsi l'influence possible de la contextualisation de *son* et empêchons que le genre soit transféré par la voie (c). À la différence de u_1, le biais intrinsèque de *son* est pris en compte. La mesure $r(u_2)$[17] quantifie alors l'effet cumulé du transfert par la voie (a), la voie (b) et l'influence lexicale directe de *son*.

La troisième modification (contexte u_3) a pour objectif de rendre les représentations du groupe sujet indifférentes au genre du nom . Le biais ne peut alors plus être dû qu'à des effets indirects (voies (b) ou (c)). Pour cela, nous considérons, comme précédemment, les paires (forme féminine, forme masculine) formées à partir d'un nom de métier donné et traduisons celles-ci après avoir remplacé, en sortie de l'encodeur, la représentation du nom de métier par la moyenne de la représentation construite avec un contexte féminin et de celle construite avec un contexte masculin[18]. Notons que, l'intervention étant réalisée en sortie de l'encodeur, les informations de genre peuvent toujours être présentes dans les représentations contextualisées des autres mots de la source. En faisant ce changement dès la première couche de l'encodeur, il serait possible d'empêcher la diffusion du genre en source et, d'une certaine manière, de rendre tous les noms de métier épicènes. Notons que comme pour le contexte u_2, cette intervention n'a pas d'effet pour les phrases dans lesquelles le genre n'est pas marqué.

6.2.2. *Résultats et analyses*

L'effet des interventions décrites ci-dessus est représenté à la figure 4. Ces résultats montrent que la neutralisation de *son* (intervention u_1) a un impact très fort sur les deux mesures que nous étudions : la figure (b) montre que la valeur moyenne[19] de $r(u)$ passe de $0{,}49$ à $0{,}18$ et sa médiane de $0{,}56$ à $0{,}34$ après modification du contexte. Modifier le contexte en gommant l'influence purement lexicale de *son* amplifie fortement le déséquilibre entre les deux pronoms puisque le $r(u)$ moyen s'éloigne de 1 et que le $b(u)$ moyen s'éloigne de 0 et, plus généralement, toute la distribution de ces deux quantités est déplacée vers les valeurs négatives. Dans ce contexte, il est permis de penser que les deux alternatives deviennent toutefois moins probables que d'autres déterminants (notamment *the*). La prédiction du déterminant anglais dépend donc fortement de la présence ou non du possessif *son* dans la source française. En revanche, une fois supprimée l'information lexicale liée à ce mot, l'influence indirecte du genre sujet dans la prédiction du pronom anglais se trouve réduite (voie (c)).

17. $b(u_2)$ et $b(u)$ sont identiques, la modification sur la couche de sortie de l'encodeur n'ayant aucun impact pour les phrases dans lesquelles le genre n'est pas marqué.

18. Cette intervention ne peut être effectuée que lorsque les segmentations de la forme féminine et de la forme masculine d'un nom de métier contiennent le même nombre d'unités sous-lexicales. Nous remplaçons alors, dans les deux phrases, la représentation de chaque unité par la moyenne des représentations féminines et masculines, position par position.

19. Les valeurs des moyennes et des médianes ont été calculées après avoir supprimé sept points aberrants qui accentuaient plus encore la baisse observée.

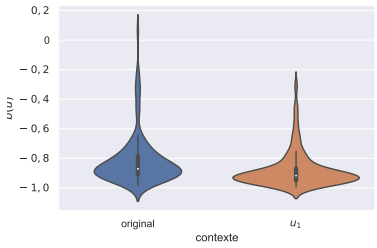
Les deux autres interventions que nous proposons ont un impact nettement plus faible sur $r(u)$: sa valeur moyenne (resp. médiane) passe de $0{,}46$ à $0{,}48$ (resp. de $0{,}558$ à $0{,}559$) après la modification du contexte u_2 et à $0{,}50$ (resp. $0{,}549$) après la modification u_3. Ces résultats agrégés masquent le fait que les interventions que nous proposons ont un fort impact sur quelques phrases pour lesquelles les valeurs de $r(u)$ sont anormalement faibles : les valeurs moyenne et médiane de $r(u)$ sont sensiblement différentes et surtout ne varient pas toujours dans le même sens (l'intervention u_3 entraîne une augmentation de la moyenne $r(u)$, mais une baisse de sa médiane).

Si l'on considère seulement l'évolution des valeurs médianes (pour limiter l'effet des valeurs extrêmes), il apparaît que la neutralisation du nom de métier (contexte u_3) entraîne une diminution faible de l'influence du genre du groupe sujet : quand l'information de genre n'est pas marquée sur le nom, la probabilité de générer *his* dans un contexte féminin et celle de le générer dans un contexte masculin deviennent un peu plus proches, avec toujours une très forte préférence pour le masculin. Ceci met en évidence l'existence de la voie (b), qui contribue toutefois peu à la sélection du genre du pronom. Gommer l'information de genre encodée dans la représentation de *son* (intervention u_2) n'a qu'un effet marginal, ce qui suggère de nouveau que la voie (c) ne joue qu'un petit rôle dans le transfert de l'information.

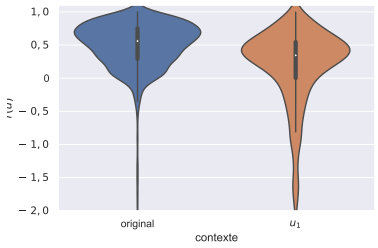
L'interprétation la plus réaliste de ces résultats est alors que l'essentiel du déséquilibre entre *his* et *her* s'explique par les corrélations entre mots anglais, capturées par l'auto-attention du décodeur : peu importe le genre du GN français, l'association entre nom et genre du pronom semble principalement due à la distribution inégalitaire observée dans la partie cible du corpus d'apprentissage. Cette interprétation, à confirmer par d'autres manipulations, dessine deux pistes non mutuellement exclusives pour réduire les biais observés : d'une part, rééquilibrer les statistiques du corpus d'apprentissage (sans nécessairement chercher à les associer à la réalité du genre correct en français, qui compte pour peu), d'autre part, renforcer explicitement les mécanismes de transfert cross-langue (voies (a) et (b)), par exemple en intégrant les dépendances syntaxiques dans le mécanisme d'attention intra ou inter-langue.

7. Conclusions

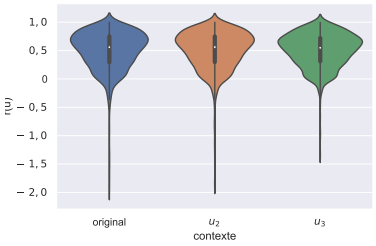
Nous avons présenté dans ce travail un nouveau jeu de tests permettant de mettre en évidence les biais de genre dans les systèmes de traduction automatique neuronale. Ce jeu de tests offre de nombreuses possibilités pour analyser finement les échanges d'informations entre les différentes composantes du réseau de neurones. En particulier, nous avons pu mettre en évidence que l'information de genre était distribuée dans les représentations de l'ensemble des tokens de la phrase source et de la phrase cible, et ce, même lorsque le patron générant les phrases fait intervenir des dépendances complexes. Nous avons également montré, grâce à des expériences consistant à intervenir sur les représentations internes du réseau de neurones, que le transfert de l'information de genre entre le français et l'anglais était complexe et reposait sur de multiples facteurs, même si le contexte de la phrase source semble jouer un rôle prépondérant.



(a) Phrases avec genre indéterminé



(b) Phrases avec genre non ambigu



(c) Phrases avec genre non ambigu

Figure 4. *Impact des interventions décrites dans la section 6.2 sur la distribution des deux mesures de biais de genre. La valeur moyenne des différentes mesures est représentée par un point blanc. Pour rendre le graphe plus lisible, les sept plus petites valeurs de $r(u)$ pour le contexte u_1 ont été supprimées de la figure (b) et les trois plus petites valeurs pour le contexte original de la figure (c).*

Pour prolonger ce travail, nous souhaitons confirmer nos observations expérimentales, notamment en étudiant de nouvelles interventions visant à découpler le décodeur de l'encodeur et à intégrer d'autres informations lexicales, comme la fréquence d'apparition des tokens étudiés. Nous espérons également parvenir à mieux caractériser les biais du corpus d'apprentissage et à étendre les méthodes décrites ici à d'autres langues et à d'autres phénomènes.

Remerciements

Ce travail a été partiellement financé par le projet NeuroViz soutenu par la Région Île-de-France dans le cadre d'un financement DIM RFSI 2020. Nous remercions le CNRS/TGIR HUMA-NUM et le Centre de calcul IN2P3 (Lyon - France) pour la fourniture des ressources informatiques et de traitement des données nécessaires à ce travail ainsi que les trois relecteurs pour tous leurs commentaires.

8. Bibliographie

Amsili P., Seminck O., « Schémas Winograd en français : une étude statistique et comportementale », *TALN 2017*, Orléans, France, p. 28-35, June, 2017.

Balvet A., « Métriques d'évaluation en Traduction Automatique : le sens et le style se laissent-ils mettre en équation ? », *in* T. Milliaressi (ed.), *La Traduction épistémique : entre poésie et prose*, Presses Universitaires du Septentrion, p. 315-356, 2020.

Basta C., Costa-jussà M. R., Fonollosa J. A. R., « Towards Mitigating Gender Bias in a decoder-based Neural Machine Translation model by Adding Contextual Information », *Proc. of the The Fourth Widening Natural Language Processing Workshop*, ACL, Seattle, USA, p. 99-102, July, 2020.

Belinkov Y., Glass J., « Analysis Methods in Neural Language Processing : A Survey », *TACL*, vol. 7, p. 49-72, 04, 2019.

Blodgett S. L., Barocas S., Daumé III H., Wallach H., « Language (Technology) is Power : A Critical Survey of "Bias" in NLP », *ACL*, ACL, Online, p. 5454-5476, July, 2020.

Burlot F., Yvon F., « Evaluating the morphological competence of Machine Translation Systems », *WMT*, ACL, Copenhagen, Denmark, p. 43-55, September, 2017.

Burlot F., Yvon F., « Évaluation morphologique pour la traduction automatique : adaptation au français », *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN*, p. 61-74, 2018.

Callison-Burch C., Osborne M., Koehn P., « Re-evaluating the role of BLEU in machine translation research », *EACL*, 2006.

Cho W. I., Kim J. W., Kim S. M., Kim N. S., « On Measuring Gender Bias in Translation of Gender-neutral Pronouns », *Proc. of the First Workshop on Gender Bias in Natural Language Processing*, ACL, Florence, Italy, p. 173-181, August, 2019.

Crawford K., « The Trouble with Bias », *Keynote at NeurIPS*, 2017.

Dister A., Moreau M.-L., *Mettre au féminin : guide de féminisation des noms de métier, fonction, grade ou titre*, 3e édition edn, Fédération Wallonie-Bruxelles, 2014.

Escudé Font J., Costa-jussà M. R., « Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques », *Proc. of the First Workshop on Gender Bias in Natural Language Processing*, ACL, Florence, Italy, p. 147-154, August, 2019.

Gonen H., Webster K., « Automatically Identifying Gender Issues in Machine Translation using Perturbations », *EMNLP*, ACL, Online, p. 1991-1995, November, 2020.

Hewitt J., Liang P., « Designing and Interpreting Probes with Control Tasks », *EMNLP*, ACL, Hong Kong, China, p. 2733-2743, November, 2019.

Huddleston R., Pullum G. K. *et al.*, *The Cambridge Grammar of English*, Cambridge University Press, 2002.

Isabelle P., Cherry C., Foster G., « A Challenge Set Approach to Evaluating Machine Translation », *EMNLP*, ACL, Copenhagen, Denmark, p. 2486-2496, September, 2017.

Kreutzer J., Bastings J., Riezler S., « Joey NMT : A Minimalist NMT Toolkit for Novices », *EMNLP, Demonstrations*, ACL, Hong Kong, China, p. 109-114, November, 2019.

Kuczmarski J., Johnson M., Gender-aware natural language translation, Technical report, 2018.

Kudo T., « Subword Regularization : Improving Neural Network Translation Models with Multiple Subword Candidates », *ACL*, ACL, Melbourne, Australia, p. 66-75, July, 2018.

Levesque H., Davis E., Morgenstern L., « The Winograd schema challenge », *KR*, 2012.

Lu K., Mardziel P., Wu F., Amancharla P., Datta A., « Gender bias in neural natural language processing », *Logic, Language, and Security*, Springer, p. 189-202, 2020.

Östling R., Tiedemann J., « Efficient word alignment with Markov Chain Monte Carlo », *Prague Bulletin of Mathematical Linguistics*, vol. 106, p. 125-146, October, 2016.

Ott M., Edunov S., Grangier D., Auli M., « Scaling Neural Machine Translation », *WMT*, ACL, Brussels, Belgium, p. 1-9, October, 2018.

Pearl J., « Direct and Indirect Effects », *UAI*, UAI '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 411–420, 2001.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., « Scikit-learn : Machine Learning in Python », *JMLR*, vol. 12, p. 2825-2830, 2011.

Prates M. O. R., Avelar P. H., Lamb L. C., « Assessing gender bias in machine translation : a case study with Google Translate », *Neural Computing and Applications*, vol. 32, n° 10, p. 6363-6381, 2020.

Renduchintala A., Williams A., « Investigating Failures of Automatic Translation in the Case of Unambiguous Gender », *CoRR*, 2021.

Rudinger R., Naradowsky J., Leonard B., Durme B. V., « Gender Bias in Coreference Resolution », *in* M. A. Walker, H. Ji, A. Stent (eds), *NAACL-HLT*, ACL, p. 8-14, 2018.

Saunders D., Byrne B., « Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem », *ACL*, ACL, Online, p. 7724-7736, July, 2020.

Saunders D., Sallis R., Byrne B., « Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It », *Proc. of the Second Workshop on Gender Bias in Natural Language Processing*, ACL, Barcelona, Spain (Online), p. 35-43, December, 2020.

Savoldi B., Gaido M., Bentivogli L., Negri M., Turchi M., « Gender Bias in Machine Translation », *Transactions of the Association for Computational Linguistics*, vol. 9, n° 0, p. 845-874, 2022.

Sennrich R., Haddow B., Birch A., « Controlling Politeness in Neural Machine Translation via Side Constraints », *NAACL*, ACL, San Diego, California, p. 35-40, June, 2016.

Stanovsky G., Smith N. A., Zettlemoyer L., « Evaluating Gender Bias in Machine Translation », *ACL*, ACL, Florence, Italy, p. 1679-1684, July, 2019.

Vanmassenhove E., Hardmeier C., Way A., « Getting Gender Right in Neural Machine Translation », *EMNLP*, ACL, Brussels, Belgium, p. 3003-3008, October-November, 2018.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. u., Polosukhin I., « Attention is All you Need », *NeurIPS*, p. 5998-6008, 2017.

Vig J., Gehrmann S., Belinkov Y., Qian S., Nevo D., Singer Y., Shieber S., « Investigating Gender Bias in Language Models Using Causal Mediation Analysis », *NeurIPS*, vol. 33, Curran Associates, Inc., p. 12388-12401, 2020.

Winograd T., *Language as a cognitive process : Volume 1 : Syntax*, Addison-Wesley Pub. Co., Reading, MA, 1983.

Zhao J., Wang T., Yatskar M., Ordonez V., Chang K.-W., « Gender Bias in Coreference Resolution : Evaluation and Debiasing Methods », *NAACL*, ACL, New Orleans, Louisiana, p. 15-20, June, 2018.