Biais de genre dans un système de traduction automatique neuronale : une étude préliminaire

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, François Yvon

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Biais de genre dans un système de traduction automatique neuronale : une étude préliminaire

Guillaume Wisniewski[1] Lichao Zhu[1] Nicolas Ballier[2] François Yvon[3]
(1) LLF, Univ. Paris & CNRS , 75013, Paris France
(2) CLILLAC-ARP, Univ. Paris, 750013, Paris France
(3) LISN, Univ. Paris-Saclay & CNRS , 91403, Orsay France
{guillaume.wisniewski,lichao.zhu,nicolas.ballier}@u-paris.fr,
francois.yvon@limsi.fr

RÉSUMÉ

Cet article présente les premiers résultats d'une étude en cours sur les biais de genre dans les corpus d'entraînements et dans les systèmes de traduction neuronale. Nous étudions en particulier un corpus minimal et contrôlé pour mesurer l'intensité de ces biais dans les deux directions anglais-français et français-anglais ; ce cadre contrôlé nous permet également d'analyser les représentations internes manipulées par le système pour réaliser ses prédictions lexicales, ainsi que de formuler des hypothèses sur la manière dont ce biais se distribue dans les représentations du système.

ABSTRACT

Gender Bias in Neural Translation : a preliminary study

This paper is a blueprint of a current study in the making on gender bias in French/English neural translation toolkits. We discuss previous research using probes for neural machine translation. We then study a minimal controlled corpus and use it to measure the intensity of such biases in the two translation directions (from and into English). Using a controlled experimental design also enables us to analyze the internal representations (attention matrices) of the translation system, and to formulate hypotheses regarding the way these biases are encoded within these representations.

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

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

1 Introduction

Il est largement admis (Callison-Burch *et al.*, 2006; Lo & Wu, 2011; Balvet, 2020) que les métriques automatiques classiques telles que les scores BLEU (Papineni *et al.*, 2002) ou METEOR (Banerjee & Lavie, 2005) sont inadaptées pour rendre compte des progrès observables en matière de qualité des traductions automatiques (TA) prédites par les systèmes neuronaux. Partant de ce constat, plusieurs protocoles (Isabelle *et al.*, 2017; Burlot & Yvon, 2017, 2018) ont été récemment proposés pour évaluer et diagnostiquer plus finement la traduction entre l'anglais et le français. 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 traits problématiques : (a) le fait que des erreurs de traductions sont plus fréquentes pour des énoncés qui mettent en scène des participantes de genre féminin ; (b) le fait que des traductions rendent linguistiquement explicites 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, 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.*, 2021) et les références citées), et des réponses idoines de la part des fournisseurs de technologie [1]. Ces questions sont en particulier discutées dans les actes de la série d'ateliers sur les biais de genre en traitement des langues [2].

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, mis 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.

La première contribution de cet article est d'étendre les analyses conduites sur la traduction depuis l'anglais vers le français dans la direction inverse, en proposant de nouveaux contrastes pour mettre en évidence et quantifier ces biais de genre. Nous nous intéressons, dans un second temps, à partir de l'analyse des représentations internes d'un système de traduction neuronale à base de TRANSFORMER, à identifier plus finement la manière dont ces biais sont encodés dans les paramètres du réseau, en particulier ceux qui servent aux calculs des matrices d'attention.

2 Un jeu de tests contrôlé pour observer les biais de genre

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

2.1 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 [3] (Rudinger *et al.*, 2018) et WinoBias [4] (Zhao *et al.*, 2018), tous deux inspirés des schémas Winograd

1. À titre d'illustration, les efforts de Google pour remédier à ces effets sont décrits dans ce billet https://ai.googleblog.com/2020/04/a-scalable-approach-to-reducing-gender.html.
2. Gender bias in NLP, http://genderbiasnlp.talp.cat.
3. https://github.com/rudinger/winogender-schemas
4. https://www.aclweb.org/anthology/attachments/N18-2003.Datasets.zip

(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 dans le prédicat induit un changement dans l'interprétation de la coréférence dans la subordonnée, qui renvoie au sujet ou à l'objet de la 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 conduit à interpréter *they* comme référant à *The city councilmen*, alors que la seconde 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.

Rudinger *et al.* (2018) décalquent ce schéma d'alternance pour 120 couples de phrases en mobilisant deux types de constructions pour constituer le corpus **Winogender** :
— *[entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances].*
— *[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/she* et son antécédent dans des phrases comme (le référent attendu est entre crochets) :

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

(3) *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 pronom/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.

2.2 Une évaluation des biais de genre

(Stanovsky *et al.*, 2019) utilise les 3 880 phrases issues de ces travaux pour mesurer les biais de systèmes traduisant depuis l'anglais vers 8 langues dans lesquelles le genre est grammaticalisé.

L'exemple (2) 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'erreurs 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. Outre son utilisation d'un corpus artificiel, on note qu'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 lequel le genre du groupe nominal sujet ne peut être déterminé :

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

(4) *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 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 le point de vue de l'énonciateur *(de dicto)* soit celui de l'actant (*de re*) :

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

(6) *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 (6), *he* pourrait renvoyer à l'objet *(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 direction français-anglais. Nos premières tentatives pour construire un jeu de test en post-éditant des traductions automatiques de WinoGender se sont rapidement heurtées à de nombreux cas d'ambiguïté 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 exploiter les matrices d'attention calculées pendant la traduction. Nous avons préféré utiliser un jeu de données plus simple dans nos expériences, en nous inspirant des travaux de Saunders & Byrne (2020) qui sont présentés à la section 5.2.

2.3 Une évaluation plus contrôlée du biais de genre

À l'instar de Saunders & Byrne (2020), nous avons construit un ensemble équilibré de 388 phrases en anglais sur le patron *The [noun] completed [his/her] work*, où [noun] est un nom de profession. Dans ces phrases, la seule marque de genre est alors portée par le pronom [7] *her/his*.

Chaque patron est instancié une fois au masculin et une fois au féminin. La mesure des performances repose sur le calcul du genre du GN traduisant [the noun] en français et 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 le nom (*l'actrice*), seulement l'article (*la juge*), ou complètement ambigu (*l'analyste*). Contrairement aux données de Stanovsky *et al.* (2019), évaluer la correction des traductions est ici facile, car la position des mots portant l'information de genre est toujours la même.

Nous avons également traduit automatiquement ces phrases en français puis corrigé / normalisé les traductions pour construire le test (vers l'anglais) sur le modèle *[det] [nom] a fini son travail*. Les noms de profession en français ont été vérifiés à partir des listes de référence (Becquer *et al.*, 1999;

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.

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 / itself*

Dister & Moreau, 2014). La vérification en anglais de la traduction correcte s'appuie sur le simple repérage du pronom (*her*/*his*) dans la phrase cible. Dans cette direction, le genre sera soit déduit de celui du groupe nominal sujet en français (inférence du genre nom ou du déterminant), soit reflétera une préférence dans l'association [nom] / [genre du pronom] en anglais.

Une version alternative de ces tests remplace [the] (en anglais) par *each* et [det] en français par l'épicène *chaque*, ceci afin qu'en français la seule marque de genre soit (éventuellement) sur le nom, ce qui simplifie l'analyse des matrices d'attention qui ont alors une forme encore plus régulière. L'ensemble des corpus ainsi construits est librement téléchargeable à partir de l'URL : https://github.com/neuroviz/neuroviz/tree/main/gender_analysis_in_mt.

3 Expérimentations et résultats globaux

3.1 Le 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 6 couches, chacune avec 8 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, 76 596 736 paramètres. Le système a été entraîné avec les données de la tâche « News » de la campagne WMT'15 [8]. Les corpus `Europarl`, `NewsCommentary` et `CommonCrawl` sont utilisés pour l'apprentissage, regroupant 4 813 682 phrases et près de 141 millions de mots 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) ; 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 (resp. 32,7) pour la direction français-anglais (resp. anglais-français).

Un autre point de comparaison est donné dans le Tableau 1, qui reproduit pour ce système les mesures de biais de genre de (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. [9] Il apparaît que notre implémentation de JoeyNMT délivre 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.

3.2 Évaluation de la traduction du genre

Nous évaluons la capacité d'un système à prédire le genre des métiers à partir du corpus décrit § 2. Pour la traduction vers l'anglais, cette évaluation est simple et repose sur la vérification du genre du pronom *her*/*his*. Vers le français, le genre du groupe nominal peut être marqué soit par le déterminant,

8. Il s'agit de la dernière campagne d'évaluation sur la paire anglais-français organisée dans le cadre de la conférence WMT (http://statmt.org/wmt15).

9. 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.

JoeyNMT		Fairseq		Systran	
Acc	ΔG_s	Acc	ΔG_s	Acc	ΔG_s
45,6	30,1	48,0	4,4	43,4	41,8

TABLE 1 – Évaluation de notre système de traduction sur les phrases de WinoGender. Pour chaque système, nous calculons l'exactitude (*accuracy*) de la prédiction du pronom, ainsi que la différence de scores F1 entre la prédiction des phrases pour les genres masculin et féminin.

soit par le nom. Sauf mention contraire, nous considérerons que le genre du GN est correctement prédit lorsque le genre du déterminant *et* le genre du nom sont tous deux corrects.

Il faut noter qu'il n'est pas toujours possible de déterminer l'information de genre dans les traductions prédites par un système de TA. En effet, dans certains cas, le système produit une traduction correcte n'utilisant pas les pronoms *her/his* (p. ex. *the programmer has finished working*) ; dans d'autres cas la traduction est complètement fausse (p. ex. « l'inspectrice a fini son travail. » a été traduit en « *the young man bent on to work.* ») ou le déterminant est traduit par *its* (53 phrases correspondant pour la plupart à des situations où le nom de métier n'a pas été traduit correctement).

3.3 Résultats expérimentaux

Appliqué au corpus décrit à la section 2, dans le sens français-anglais, notre système prédit correctement le genre du pronom possessif anglais dans 65,7% des cas. Dans le sens anglais-français, il traduit correctement le genre du GN dans 46,1% des cas (respectivement 60,5% des cas pour le genre du nom, et 55,6% pour le genre du déterminant). Ces résultats [10] médiocres montrent qu'un système neuronal «standard» a des difficultés à modéliser et à prédire les informations de genre au cours de la traduction. Pour les mettre en perspective, nous avons également utilisé, pour traduire ces mêmes corpus, un moteur « grand public » `DeepL` (version 1.12.0) et `e-translation` [11], moteur de traduction développé par la Commission Européenne et librement accessible à des fins de recherche académique. Si les résultats de DeepL sont meilleurs (voir le tableau 2), ils restent très imparfaits et montrent que la tâche proposée est difficile. À notre grande surprise, les traductions de `e-translation` ne font que très peu d'erreur dans la traduction du genre, ce qui laisse supposer que ce système intègre un traitement spécifique de ces phénomènes.

Les résultats détaillés sont dans le tableau 2. Ils montrent que les systèmes considérés présentent des taux d'erreurs très différents entre genres, avec des écarts marqués entre systèmes et directions de traduction. Cette observation justifie d'étudier simultanément les deux directions. Pour notre système de traduction, la plupart des erreurs de prédiction portent sur le féminin : par exemple, pour la traduction vers l'anglais, le taux d'erreur pour les pronoms féminins est de 70,67% contre 2,87% pour les pronoms masculins.

À l'inverse, DeepL privilégie quasi systématiquement le féminin lorsqu'il traduit vers l'anglais : 252 des prédictions du système contiennent *her*, quand seulement 141 contiennent *his* (dans de rares cas, le système propose également l'alternative *his or her*). Ce comportement est inversé pour la

10. Ces calculs ignorent les hypothèses de traduction pour lesquelles aucun genre ne peut être déterminé. Pour le système français-anglais, cela arrive pour 16% des sorties, qui ne contiennent ni *her* ni *his*. Pour le système anglais-français, le genre du GN sujet n'a pu être déterminé dans 17% des cas

11. ec.europa.eu/cefdigital/eTranslation

dét. (fr)	nom (fr)	fr → en			en → fr		
		JoeyNMT	DeepL	e-translation	JoeyNMT	DeepL	e-translation
l'	épicène	46,3	53,2	68,8	77,3	85,4	84,4
	féminin	16,0	93,3	100	4,3	10,0	21,4
	masculin	90,9	54,0	100	67,7	94,6	86,1
la	épicène	26,4	94,9	100	0,0	15,4	58,3
	féminin	62,8	96,5	98,1	2,0	22,0	37,3
le	épicène	95,1	65,9	100	87,0	95,7	88,6
	masculin	100,0	70,1	100	82,0	98,9	92,5

TABLE 2 – Pourcentage de succès dans le transfert du genre entre français et anglais. La 2ème colonne distingue les cas où le nom de métier est genré en français et la première indique le déterminant du nom de métier (ces valeurs sont déterminées sur la référence pour la traduction de l'anglais vers le français).

traduction vers le français : dans ce cas, les noms de métiers sont presque toujours traduits par un masculin.

4 Analyses de la propagation de l'information de genre

Dans cette section, nous présentons plusieurs analyses complémentaires portant sur la manière dont l'information de genre est propagée depuis le GN en français vers le pronom anglais. Notre principale objectif est de déterminer quelles sont les éléments mis en jeu dans le choix du pronom *his*/*her*. En particulier, nous nous intéressons à la manière dont les représentations et les différents scores d'attention sont influencés par les informations de genre.

Rappelons que dans une architecture transformer standard, trois mécanismes attentionnels sont simultanément à l'œuvre : l'auto-attention de l'encodeur, qui permet que les représentations des tokens sources s'influencent mutuellement ; l'auto-attention du décodeur, qui joue un rôle similaire côté cible, sous la contrainte que chaque mot n'a accès qu'aux représentations des mots qui le précèdent ; enfin l'attention croisée source-cible dans le décodeur, qui permet de contextualiser les représentations cibles en les combinant avec les représentations source sur la dernière couche de l'encodeur. Nous nous intéressons ici principalement à l'auto-attention de l'encodeur.

4.1 Impact du déterminant du GN

Pour mesurer l'impact du déterminant du GN sujet, nous avons réalisé une première expérience de contrôle en construisant un nouveau corpus de test identique à celui construit à la section 2 mais dans lequel nous neutralisons tous les déterminants dont la forme varie avec le genre : dans ce corpus, les déterminants des noms de métiers ont été systématiquement remplacés par le déterminant épicène *chaque*. Le genre à transférer n'est alors marqué que sur le nom de métier et, donc indéterminé quand ce dernier est épicène. Dans cette configuration, notre système ne commet aucune erreur en transférant en anglais le genre du nom de métier masculin, alors qu'il se trompe presque systématiquement (94,55 % d'erreurs) pour les féminins.

4.2 Le genre de *son*

La question que nous étudions dans cette section porte sur le transfert de l'information de genre entre langues. Pour traduire correctement le genre du GN français, trois hypothèses (non mutuellement exclusives) sont envisagées : (a) une influence *directe* par le calcul de l'attention cross-lingue effectuée lors de la traduction du pronom ; (b) une influence *indirecte* passant par l'encodage (cross-lingue) 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 du possessif français *son*, qui est ensuite propagée (cross-lingue) vers le pronom anglais. Ces trois possibilités sont résumées dans la figure 1. Nous nous intéressons à valider ou invalider l'existence du mécanisme (c), en analysant de plusieurs manières la représentation de *son*. Ce choix est motivé par la structure systématique des phrases françaises, qui ont recours au même mot, qui de surcroît se trouve à la même position, et dont la représentation est donc facile à extraire et manipuler.

Sonder *son* La première méthode repose sur l'utilisation de sondes linguistiques (*probes*) (Belinkov & Glass, 2019) et consiste à tester la capacité de prédire le genre du GN en observant seulement la représentation du mot *son* construite par le système de TA. En utilisant l'encodeur du système de traduction, nous calculons le vecteur de représentation associé à ce mot pour les 388 phrases de notre corpus. Nous entraînons ensuite un classifieur linéaire simple qui doit prédire le genre du GN à partir du seul vecteur représentant le possessif : l'hypothèse est que s'il est possible de réaliser avec succès cette prédiction, c'est que les représentations de *son* pour les phrases comportant un GN masculin diffèrent de celles qui comportent un GN féminin, et pourront donc influencer utilement le choix du pronom en anglais.

Expérimentalement, nous apprenons un modèle de régression logistique avec `scikit-learn` (Pedregosa *et al.*, 2011) en utilisant 75% des données, et calculons les taux d'erreurs sur les 25% restant. Cette expérience est répétée 100 fois pour pouvoir calculer l'intervalle de confiance de la précision. Compte tenu du rapport entre la taille des vecteurs d'entrée (512) et le nombre de données d'apprentissage (388), il est nécessaire de régulariser fortement ce classifieur, ce que nous obtenons en ajoutant à la fonction objectif une pénalité ℓ_1 [12]. D'une manière générale, il importe de contrôler la capacité des sondes, et de s'assurer qu'elles ne pourraient pas également apprendre des étiquetages aléatoires (Hewitt & Liang, 2019).

Les résultats sont rapportés dans le tableau 3. On constate que l'information de genre est effectivement présente, mais uniquement dans les couches les plus profondes : la précision du classifieur utilisant comme caractéristique les représentations de `son` issues des deux premières couches est très faible (proche de celle d'un classifieur prenant ses décisions au hasard) mais augmente rapidement (de plus de 20 points entre la 2ème et la 4ème couche) pour se stabiliser autour de 80%. Suivant les recommandations de (Hewitt & Liang, 2019) nous avons également calculé la précision de notre sonde après avoir appliqué une permutation aléatoire des étiquettes afin de nous assurer que la sonde ne capturait pas des corrélations fallacieuses. Les résultats rapportés à la table 3 montre que les informations de genre sont bien présentes dans la représentation et non dans la sonde.

Manipuler *son* La seconde méthode que nous proposons utilise *une intervention* pour mettre en évidence ce même effet. Elle consiste à remplacer le vecteur représentant ce mot à la sortie de

12. Au final, dans nos expériences entre 30% et 80% des paramètres sont nuls.

Couche	Précision sonde	Précision aléatoire
1	$57,4\%$ ± 0.8	$45,3\%$ ± 0.9
2	$60,4\%$ ± 1.0	$50,7\%$ ± 0.8
3	$72,5\%$ $\pm 0,8$	$48,8\%$ ± 0.9
4	$82,0\%$ $\pm 0,6$	$48,6\%$ ± 0.8
5	$81,9\%$ $\pm 0,7$	$49,6\%$ ± 0.8
6	$79,3\%$ $\pm 0,7$	$49,2\%$ ± 0.8

TABLE 3 – Précision d'un classifieur prédisant le genre du GN à partir de la représentation de `son`.

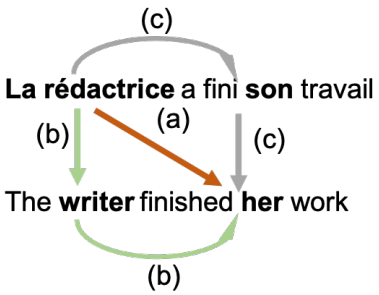

FIGURE 1 – Les différents éléments pouvant influencer le choix du genre du pronom possessif en anglais.

l'encodeur alternativement par (a) une version *neutre*, obtenue en moyennant les représentations de *son* sur l'ensemble des phrases (notre corpus est construit avec la contrainte que le nombre de GN féminin est le même que le nombre de GN masculin) ; (b) une représentation *masculine* (resp. *féminine*) supposée prototypique, obtenues en encodant respectivement les deux phrases suivantes :

(7) le facteur a terminé son travail.

(8) la pharmacienne a terminé son travail.

Ces deux phrases ont été choisies parce que, dans les deux cas, la traduction du genre par notre système est correcte et parce que l'information du genre est portée à la fois par le déterminant et par le nom. Le reste de la traduction se déroule sans autre modification, ce qui nous permet de comparer l'influence de ces 4 représentations (trois fixes, une contextuelle) sur la traduction du pronom.

Les résultats figurent dans le tableau 4 : nous y avons rapporté la proportion d'hypothèses de traduction dans lesquelles le pronom possessif est féminin, masculin ou ne peut pas être déterminé, en fonction de l'intervention sur la représentation de `son`. Contrairement à ce qui était attendu, modifier la représentation du pronom possessif n'a que peu d'impact sur le choix du pronom possessif *his* ou `her`. Ces résultats montrent que la représentation de `son` construite par le système de TA n'est pas (ou peu) utilisée lors de la génération de l'hypothèse de traduction, bien que les résultats rapportés dans le paragraphe précédent montrent que celles-ci sont particulièrement pertinentes. Ce résultat contre intuitif rejoint plusieurs observations dans la littérature des sondes : ce n'est pas parce qu'une information « linguistique » est encodée dans les représentations neuronales qu'elle est exploitée par le réseau (Belinkov & Glass, 2019).

4.3 Analyse des matrices d'attention

Un des intérêts du corpus proposé dans ce travail est de faciliter l'analyse des matrices d'attention au cœur des systèmes de traduction neuronaux. En effet, les phrases ayant une structure fixe, il est aisé d'identifier la position des mots marquant le genre (p. ex. pour la traduction depuis le français du déterminant et du nom de métier) et d'analyser l'attention depuis ou vers ces positions. Nous présentons quelques résultats préliminaires illustrant le type d'études réalisables à partir de ce corpus.

intervention	genre de la traduction	% erreur	intervention	genre de la traduction	% erreur
aucune	féminin	10,6%	moyenne	féminin	10,4%
	indéterminé	13,4%		indéterminé	13,0%
	masculin	76,0%		masculin	76,6%
féminin	féminin	12,5%	masculin	féminin	11,2%
	indéterminé	10,6%		indéterminé	14,0%
	masculin	76,9%		masculin	74,8%

TABLE 4 – Manipulation des représentations de *son* : proportion d'hypothèses de traduction dans lesquelles le pronom possessif est féminin, masculin ou ne peut pas être déterminé, en fonction de l'intervention sur la représentation de `son`.

	couche n° 0		couche n° 1		couche n° 2		couche n° 3		couche n° 4		couche n° 5	
↓ préd. / vers →	dét.	mét.	dét.	mét.	dét.	mét.	dét.	mét.	dét.	mét.	dét.	mét.
correcte	0.018	0.111	0.069	0.134	0.178	0.114	0.174	0.072	0.147	0.151	0.228	0.203
incorrecte	0.017	0.074	0.063	0.137	0.173	0.114	0.181	0.081	0.153	0.140	0.226	0.211

TABLE 5 – Score d'auto-attention moyen entre le possessif *son* et le déterminant du nom de métier (dét.) ou le nom de métier (mét.), suivant que le genre soit prédit correctement ou non.

Pour mieux comprendre les informations utilisées pour prédire le genre du pronom anglais, nous représentons dans le tableau 5, la moyenne sur le corpus du score d'auto-association entre « *son* » et les deux mots du GN. Le modèle ayant 8 têtes d'attention par couches, nous avons considéré, pour chaque phrase, uniquement le plus grand score, ce qui revient à vérifier qu'au moins une tête pointe sur un des mots permettant de prédire le genre. Lorsque le nom de métier est segmenté en plusieurs unités sous-lexicales, seul le plus grand score d'attention vers une de ces unités est conservé."

Les résultats du tableau 5 montrent que les deux positions permettant de traduire correctement le mot « *son* » sont utilisées pour construire la représentation de ce mot : dans les deux dernières couches, le score d'attention entre «*son*» et ces deux mots dépasse 0,2 [13]. On note que ces scores sont comparables dans les traductions « correctes » et « erronées », à l'exception de l'attention vers le nom de métier sur la première couche. Ceci confirme les résultats présentés supra : le modèle s'appuie plus sur le nom de métier pour prendre ses décisions que sur l'article.

5 Travaux connexes : mesurer et corriger les biais de genre

5.1 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). 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

13. Les scores d'auto-attentions sont positifs et normalisés de façon que toutes les auto-attentions entre un mot et les autres mots de la phrase somment à 1 ; dans la mesure où les phrases du corpus sont constitués en moyenne de 7 mots et que l'auto-attention entre un mot et lui-même est toujours élevée, une valeur de 0,2 peut être considérée comme importante.

(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 [14] 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 [15]. 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 [16] est explicite dans le texte source, auquel cas il est attendu qu'elle soit transférée correctement dans le texte cible, afin toujours de préserver les intérêts de l'auteur ainsi que celui des personnes qui seraient évoquées dans le texte. De nouveau, 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 — ce qui est possible quand le même genre est marqué sur plusieurs éléments du discours). En particulier entre dans cette catégorie le fait ne pas préserver l'ambiguïté ou la fluidité du genre alors que des pronoms sont disponibles pour éviter des assignations de genre binaire (voir pour l'anglais l'article de synthèse de (Cao & Daumé III, 2019)).

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 & Johnson, 2018; Prates *et al.*, 2019), en étudiant aussi bien des corpus artificiels que des corpus réels (Gonen & Webster, 2020). Notons que la question du genre en TA peut être abordée sous d'autres angles : ainsi, Vanmassenhove *et al.* (2018) présente 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 10 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.

14. 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.

15. 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.

16. 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.

5.2 Atténuer automatiquement les biais de genre

Mesurer les biais permet aussi d'évaluer l'impact de travaux visant à les atténuer dans des traductions automatiques. Ces travaux mobilisent principalement trois types de techniques (voir (Savoldi *et al.*, 2021) pour une étude récente). Une première consiste à manipuler les représentations lexicales. Elle est utilisée par Escudé Font & Costa-jussà (2019) qui injectent dans le système OpenNMT des plongements lexicaux entraînés avec l'algorithme *gender-neutral GloVe* de Zhao *et al.* (2018). Ils 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 pré-annotation (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'indiciation 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 direction anglais-espagnol et des analyses complémentaires sont réalisées par Saunders *et al.* (2020). Cette technique est enfin utilisée par Kuczmarski & 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'approches 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 & Byrne (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 genres 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).

6 Perspectives et Conclusions

Nous avons introduit dans ce travail un nouveau jeu de test permettant de mettre en évidence les biais de genre dans les systèmes de traduction automatique. Ce jeu de tests offre de nombreuses possibilités pour analyser finement les échanges d'informations entre les différentes composantes du réseau de neurones que nous souhaitons explorer dans nos travaux futurs. Nous pensons en effet, qu'en plus de leur quantification, une meilleure compréhension des causes des biais sont une étape nécessaire à la « neutralisation » de ceux-ci.

Remerciements

Ce travail a été partiellement financé par le projet NeuroViz / Explorations et visualisations d'un système de traduction neuronale, soutenu par la Région Ile-de-France dans le cadre d'un financement DIM RFSI 2020.

Références

AMSILI P. & SEMINCK O. (2017). Schémas Winograd en français : une étude statistique et comportementale. In *TALN 2017*, p. 28–35, Orléans, France. HAL : hal-01628342.

BALVET A. (2020). Métriques d'évaluation en traduction automatique : le sens et le style se laissent-ils mettre en équation ? In T. MILLIARESSI, Éd., *La Traduction épistémique : entre poésie et prose*, p. 315–356. Presses Universitaires du Septentrion.

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, p. 65–72, Ann Arbor, Michigan.

BASTA C., COSTA-JUSSÀ M. R. & FONOLLOSA J. A. R. (2020). Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, p. 99–102, Seattle, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2020.winlp-1.25.

BECQUER A., CERQUIGLINI B., CHOLEWKA N., COUTIER M., FRÉCHER J. & MATHIEU M.-J. (1999). *Femme, j'écris ton nom...* La Documentation française.

BELINKOV Y. & GLASS J. (2019). Analysis Methods in Neural Language Processing : A Survey. *Transactions of the Association for Computational Linguistics*, **7**, 49–72. DOI : 10.1162/tacl_a_00254.

BLODGETT S. L., BAROCAS S., DAUMÉ III H. & WALLACH H. (2020). Language (technology) is power : A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5454–5476, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.485.

BURLOT F. & YVON F. (2017). Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, p. 43–55, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-4705.

BURLOT F. & YVON F. (2018). Évaluation morphologique pour la traduction automatique : adaptation au français. In *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN*, p. 61–74.

CALLISON-BURCH C., OSBORNE M. & KOEHN P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

CAO Y. T. & DAUMÉ III H. (2019). Toward gender-inclusive coreference resolution. arXiv preprint http://arxiv.org/abs/1910.13913.

CHO W. I., KIM J. W., KIM S. M. & KIM N. S. (2019). On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, p. 173–181, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-3824.

DISTER A. & MOREAU M.-L. (2014). *Mettre au féminin : guide de féminisation des noms de métier, fonction, grade ou titre*. Fédération Wallonie-Bruxelles, 3e édition édition.

ESCUDÉ FONT J. & COSTA-JUSSÀ M. R. (2019). Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, p. 147–154, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-3821.

GONEN H. & WEBSTER K. (2020). Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1991–1995, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.180.

HEWITT J. & LIANG P. (2019). Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 2733–2743, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1275.

HUDDLESTON R., PULLUM G. K. *et al.* (2002). *The Cambridge Grammar of English*. Cambridge University Press.

ISABELLE P., CHERRY C. & FOSTER G. (2017). A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2486–2496, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1263.

KREUTZER J., BASTINGS J. & RIEZLER S. (2019). Joey NMT : A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) : System Demonstrations*, p. 109–114, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-3019.

KUCZMARSKI J. & JOHNSON M. (2018). Gender-aware natural language translation. *Technical Disclosure Commons*, p. 1–9.

KUDO T. (2018). Subword regularization : Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 66–75, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1007.

LEVESQUE H., DAVIS E. & MORGENSTERN L. (2012). The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

LO C.-K. & WU D. (2011). MEANT : An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 220–229.

LU K., MARDZIEL P., WU F., AMANCHARLA P. & DATTA A. (2020). Gender bias in neural natural language processing. In *Logic, Language, and Security*, p. 189–202. Springer. DOI : 10.1007/978-3-030-62077-6_14.

OTT M., EDUNOV S., GRANGIER D. & AULI M. (2018). Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 1–9, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6301.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

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. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PRATES M. O., AVELAR P. H. & LAMB L. C. (2019). Assessing gender bias in machine translation : a case study with Google translate. *Neural Computing and Applications*, p. 1–19.

RUDINGER R., NARADOWSKY J., LEONARD B. & DURME B. V. (2018). Gender bias in coreference resolution. In M. A. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, p. 8–14 : Association for Computational Linguistics. DOI : 10.18653/v1/n18-2002.

SAUNDERS D. & BYRNE B. (2020). Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7724–7736, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.690.

SAUNDERS D., SALLIS R. & BYRNE B. (2020). Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, p. 35–43, Barcelona, Spain (Online) : Association for Computational Linguistics.

SAVOLDI B., GAIDO M., BENTIVOGLI L., NEGRI M. & TURCHI M. (2021). Gender bias in machine translation. arxiv preprint http://arxiv.org/abs/2104.06001.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 35–40, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1005.

STANOVSKY G., SMITH N. A. & ZETTLEMOYER L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1679–1684, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1164.

VANMASSENHOVE E., HARDMEIER C. & WAY A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3003–3008, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1334.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30*, p. 5998–6008 : Curran Associates, Inc.

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

ZHAO J., WANG T., YATSKAR M., ORDONEZ V. & CHANG K.-W. (2018). Gender bias in coreference resolution : Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 15–20, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2003.