Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Flux d'informations dans les systèmes encodeur-décodeur. Application à l'explication des biais de genre dans les systèmes de traduction automatique.

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

RÉSUMÉ
Ce travail présente deux séries d'expériences visant à identifier les flux d'information dans les systèmes de traduction neuronaux. La première série s'appuie sur une comparaison des décisions d'un modèle de langue et d'un modèle de traduction pour mettre en évidence le flux d'information provenant de la source. La seconde série met en évidence l'impact de ces flux sur l'apprentissage du système dans le cas particulier du transfert de l'information de genre.

ABSTRACT
Information flow in encoder-decoder systems applied to the explanation of gender bias in machine translation systems.

This work describes two series of experiments designed to identify information flows in neural translation systems. First, we compare the decisions of a language model and of a translation model to highlight the information flow from source to target sentences. We then evaluate the impact of these flows on training results in the particular case of gender transfer.

MOTS-CLÉS : Traduction Automatique Neuronale, Explicabilité, Biais de Genre.

KEYWORDS: Neural Machine Translation, Explicability, Gender Biais.

1 Introduction

Ce travail a pour objectif d'étudier les mécanismes internes des systèmes de traduction neuronaux et notamment les flux d'informations entre l'encodeur et le décodeur : notre objectif est, à terme, d'*expliquer* les choix du décodeur en distinguant les informations qui sont capturées dans la source et transférées vers la cible, de celles qui sont extraites uniquement d'une analyse du préfixe cible. Une meilleure compréhension de ces flux permettrait d'expliquer et de corriger les biais de genre qui émaillent les traductions engendrées par ces systèmes (Stanovsky *et al.*, 2019) et portent atteinte à l'équité *(fairness)* de ces derniers (Stanczak & Augenstein, 2021; Savoldi *et al.*, 2021). Cette étude permettra également de progresser vers une meilleure compréhension de la capacité de généralisation de ces architectures, premier pas pour développer des systèmes capables d'apprendre à partir de moins de données (Baroni, 2020).

Pour cela, nous décrivons deux séries d'expériences construites autour d'un jeu de données contrôlé introduit à la section 2. La première série (§4) repose sur une comparaison des décisions prises par un modèle de langue et un modèle de traduction réalisant un décodage forcé de la phrase cible afin de mesurer l'impact des informations provenant du contexte cible (seules informations capturées par un modèle de langue) et, par contraste avec un modèle de traduction (qui s'appuie sur les informations de la source et de la cible), celles provenant de la phrase source. Pour affiner ces observations, nous nous concentrons, dans la section 5, sur le transfert de l'information de genre et montrons que, dans de nombreux cas, le système de traduction est effectivement capable de transférer l'information de genre de la source. Enfin, à la section 6 nous évaluons l'impact de celles-ci sur l'apprentissage.

2 Un jeu de données contrôlé pour étudier les flux d'information

Pour étudier les flux d'information dans une architecture de type encodeur-décodeur, nous considérons des phrases parallèles qui suivent toutes le même motif. Ce motif a été choisi pour satisfaire deux propriétés : premièrement, il est construit autour de collocations dont la présence dans le corpus d'apprentissage est avérée. Intuitivement, ces collocations dépendent fortement (voire uniquement) d'informations présentes dans la cible et peuvent être prédites (en partie au moins) sans connaissance de la source. Deuxièmement, il met également en jeu des unités lexicales qui ne peuvent être générées qu'en analysant la phrase source. Un motif ayant cette dernière propriété a été introduit par Wisniewski *et al.* (2021) :

(1) `[DET] [N]` a terminé son travail.

(2) The `[N]` has finished `[PRO]` work.

où `[N]` est un nom de métier féminin ou masculin et `[DET]` l'article défini (*le*, *la* ou `l'`) et `[PRO]` le pronom possessif (*his* ou *her*). (Dister & Moreau, 2014) fournit une liste de 1 697 noms de métiers en français associant forme féminine et forme masculine. Cette liste a été traduite de manière semi-automatique en anglais par les auteurs de (Wisniewski *et al.*, 2021), avec pour résultat un corpus de 3 394 phrases. Ce motif s'appuie sur les différentes manières dont le genre peut être exprimé en français et en anglais : la forme du pronom possessif anglais peut, en effet, se déduire, suivant les cas, du déterminant du groupe sujet français *(la/le bibliothécaire)*, de la forme du nom de métier en français *(l'infirmière/infirmier)*, ou, plus rarement, de la forme du nom de métier en anglais *(actor/actress)*. La comparaison de la capacité d'un système de traduction à capturer et à traduire correctement le genre dans ces différents cas nous permettra de pour mettre en évidence les mécanismes de transfert de genre entre ces deux langues.

Nous avons généralisé ce motif afin d'assurer que certaines parties de la phrase cible puissent être traduites sans considérer la source (typiquement parce que faisant partie d'une collocation) et d'autres, au contraire, ne puissent être traduites qu'en extrayant les informations pertinentes en source. La première modification introduit un 4-gramme fréquent dans les données d'apprentissage, *le président Barack Obama* (154 occurrences dans le corpus d'apprentissage) qui devrait facilement être « capturé » par un modèle de langue ; la seconde repose sur l'utilisation d'un verbe à particule *(phrasal verb)*, *carry out* [1], à la place du verbe *finish*. Il est attendu que la probabilité de générer la particule *out* soit

1. Il y a 10 314 occurrences de *carried out* dans le corpus d'apprentissage, 16 de *carried out [...] his work* et 3 de *carried out [...] her work*.

principalement déterminée par la présence ou l'absence du verbe anglais dans la traduction en langue cible, indépendamment de la phrase source. Au final, nous considérons la variation suivante :

(3) le président Barack Obama a pris note que `[DET] [N]` a mené à bien son travail.

(4) President Barack Obama took note that the `[N]` has carried out `[PRO]` work.

3 Données et Méthodes

Modèle de traduction et modèle de langue Nous utilisons `JoeyNMT` (Kreutzer *et al.*, 2019), qui est une implémentation d'un système de traduction automatique (TA) basé sur l'architecture TRANSFORMER. L'encodeur et le décodeur sont composés de 6 couches et chaque couche comporte 8 têtes d'attention. Les couches *feed-forward* disposent de 2 048 paramètres et la dimension des plongements lexicaux est de 512. Notre modèle comprend un total de 76 596 736 paramètres.

Nous utilisons également notre implémentation d'un modèle de langue TRANSFORMER, en utilisant PYTORCH (Paszke *et al.*, 2019). Ce modèle possède les mêmes paramètres (nombre de têtes, nombre de couches, ...) que le décodeur du système de TA. Afin d'imiter le décodeur, nous utilisons un modèle de langue incrémental (causal) dans lequel la représentation du i^{e} token est conditionnée par les $(i-1)$ tokens précédents.

Les deux modèles sont entraînés en optimisant l'entropie croisée avec ADAM sur les mêmes données (voir ci-dessous) et obtiennent respectivement un score BLEU de 34,0 et une perplexité de 43,0 sur les données de test de WMT'14.

Corpus d'apprentissage Nous utilisons le corpus parallèle anglais-français de la tâche « News » de WMT'15 qui contient 4 813 682 phrases et environ 141 millions de mots français. Ce corpuscomporte un biais très fort envers les contextes masculins : si l'on s'intéresse, par exemple, au pronom possessif (au cœur de l'approche que nous proposons pour identifier les flux d'information), le corpus WMT'15 comporte deux fois plus d'occurrences de `his` que d'occurrences de `her` (108 364 contre 47 444) et le pronom possessif `son` y est traduit 3 fois plus souvent par `his` que par `her`.

Nous avons segmenté tout le corpus d'apprentissage et les corpus de test décrits à la section 2 en unités sous-lexicales, en utilisant le modèle *unigram* de `SentencePiece` (Kudo & Richardson, 2018) ; les vocabulaires contiennent 32 000 unités dans chaque langue.

4 Flux d'information dans une architecture encodeur-décodeur

Pour caractériser le flux d'informations dans une architecture de type encodeur-décodeur, nous comparons les prédictions d'un modèle de traduction (TM dans la suite) à celles d'un modèle de langue (LM). En effet, dans un modèle de traduction, le i^{e} mot de l'hypothèse de traduction est prédit à partir des informations extraites de la phrase source $\mathbf{s}$ et des tokens cibles $t_{<i} = t_1, ..., t_{i-1}$ correspondant au préfixe de l'hypothèse de traduction déjà généré ; dans un modèle de langue, la phrase est produite de manière incrémentale et chaque mot ne dépend que des mots qui le précèdent, ce qui correspond à l'information provenant du contexte cible pour TM. Formellement, un modèle de

traduction est un *modèle de langue conditionnel* qui calcule la probabilité $p(t_i|t_{<i}, \mathbf{s})$, quand qu'un modèle de langue calcule $p(t_i|t_{<i})$. En comparant ces deux distributions de probabilité, nous pouvons évaluer l'impact des informations provenant de la source.

Plus précisément, nous avons réalisé un décodage forcé des phrases anglaises du corpus contrôlé décrit à la section 2. Celui-ci nous a permis de calculer pour chaque position de la phrase cible la distribution $p(\cdot|\mathbf{c})$ décrivant la probabilité de générer un token cible connaissant le contexte $\mathbf{c}$. Ce contexte peut être constitué soit uniquement des tokens cibles précédents (dans le cas où la probabilité serait estimée par un modèle de langue), soit des tokens cibles précédents *et* des tokens sources (lorsque la probabilité est estimée par un modèle de traduction). Comme nous réalisons un décodage forcé, le contexte est toujours constitué par les tokens cibles « de référence », ce qui permet de neutraliser le bruit lié aux erreurs de traduction et à la variabilité de celle-ci.

Nous reportons dans la table 1, pour chaque position de la phrase cible, la moyenne et la médiane de l'entropie de la distribution $p(t|\mathbf{c})$ et la moyenne et la médiane du rang du mot correct lorsque ces probabilités sont ordonnées par ordre décroissant [2]. Notons que ces deux mesures — la moyenne et la médiane — sont identiques pour tous les tokens du préfixe lorsque le contexte ne dépend que de la phrase cible puisque toutes les phrases ont le même préfixe. L'entropie de la distribution nous permet de caractériser l'incertitude dans le choix d'un token (plus l'entropie est petite, plus la masse de probabilité est « concentrée » sur un petit nombre de tokens) ; le rang permet de caractériser la capacité du modèle à prendre une bonne décision (c.-à-d. à retrouver le token de référence). Pour faciliter la lecture des résultats, nous avons remplacé les différents tokens résultant de la segmentation des noms de métier par des noms « génériques » indiquant la position du token dans la décomposition sous-lexicale et avons agrégé les statistiques des tokens aux positions 3, 4, 5 et 6 sous un même nom. Seul le nom de métier est segmenté en plusieurs unités sous-lexicales dans ces phrases standardisées.

La table 1 montre que les décisions prises par un modèle de langue ou un modèle de traduction ne sont pas « uniformes » sur la phrase : certains mots (voire certaines parties) de la phrase cible sont choisis sans difficulté (rang faible) avec une très forte confiance (entropie faible). Ces résultats montrent également que certaines parties de la cible peuvent être générées sans connaissance de la source : l'entropie de la distribution estimée par un modèle de langue est très faible à l'intérieur des collocations que nous avons introduites (notamment pour *Obama*, *out* et *that*) montrant que la connaissance du début de la collocation permet de prédire la fin de celle-ci avec une grande confiance.

Il est également intéressant de noter que l'entropie des distributions estimées par le modèle de traduction est en moyenne 1,8 fois plus petite que celle estimée par le modèle de langue. Si cette réduction est attendue (puisque pour tout couple de variables aléatoires X et Y, $H(X) - H(X|Y) \geq 0$) elle reste toutefois particulièrement forte et, surtout, dans la majorité des cas, la prise en compte du contexte source permet de prédire correctement le mot suivant de la référence.

5 Impact sur la prédiction du genre

Nous proposons maintenant d'utiliser le principe exposé dans la section précédente pour déterminer comment les flux d'information dépendent du genre du groupe sujet français. De nombreux travaux ont mis en évidence les biais présents dans les systèmes de traduction lorsque ceux-ci traduisent des informations dépendant du genre (Stanovsky *et al.*, 2019; Wisniewski *et al.*, 2021). Les méthodes

2. Nous avons reporté la moyenne et la médiane pour détecter la présence d'éventuels points aberrants dans les distrbutions.

	entropie				rang			
	moyenne		médiane		moyenne		médiane	
	LM	TM	LM	TM	LM	TM	LM	TM
_president	5,65	1,12	5,65	1,08	172	1	172	1
_barack	3,70	0,78	3,70	0,79	7	1	7	1
_obama	0,01	1,47	0,01	1,47	1	1	1	1
_took	4,30	2,86	4,30	2,86	34	3	34	3
_note	3,46	0,90	3,46	0,90	23	1	23	1
_that	0,85	1,89	0,85	1,89	5	1,03	5	1
_the	4,46	2,27	4,46	2,10	1	1,08	1	1
noun@0	5,84	3,89	5,84	3,81	9004,92	863,44	7882,50	1
noun@1	3,76	3,03	3,93	2,92	1188,73	487,48	79	1
noun@2	3,77	2,53	4,08	2,09	1039,03	403,31	19	1
noun≥3	4,25	2,68	4,33	2,09	1234,09	477,71	80	1
_has	5,08	3,78	5,26	3,71	20,92	10,07	6	1
_carried	5,42	3,33	5,53	3,31	196,18	2,98	177,50	3
_out	1,79	2,10	1,53	2,11	1,06	1	1	1
_her	4,00	2,18	3,99	2,17	62,62	3,03	50	3
_his	3,99	1,78	3,99	1,72	3,95	1,03	3	1
_work	5,85	3,23	5,91	3,22	4,57	1	3	1
_.	3,43	3,10	3,42	3,14	214,97	1,07	201	1

TABLE 1 – Caractérisation de la probabilité de générer les mots de la référence, pour chaque position d'une phrase contrôlée, lors d'un décodage forcé par un modèle de langue et de traduction.

introduites dans la section précédente permettent d'expliquer, en partie, ces observations.

Nous proposons de comparer la capacité d'un modèle de langue et d'un modèle de traduction à prédire la forme correcte du pronom possessif en anglais dans des phrases suivant le motif (1)-(2) [3], en comparant les valeurs estimées par ces deux modèles de $p\,(\texttt{her}|c)$ et $p\,(\texttt{his}|c)$. Ces probabilités peuvent être estimées facilement en réalisant un décodage forcé de la phrase cible. Nous estimons la préférence d'un modèle à générer `her` ou `his` en calculant [4] : $b(c) = 1 - \frac{2\times p(\texttt{her}|c)}{p(\texttt{his}|c)+p(\texttt{her}|c)}$ Intuitivement, plus b est proche de -1 (resp. 1), plus $p(\texttt{her}|c)$ est grand (resp. petit) devant $p(\texttt{his}|c)$ et plus le modèle aura une préférence pour produire `her` (resp. `his`) ; b est proche de 0 quand le modèle n'a pas de préférence claire pour l'une ou l'autre des deux formes.

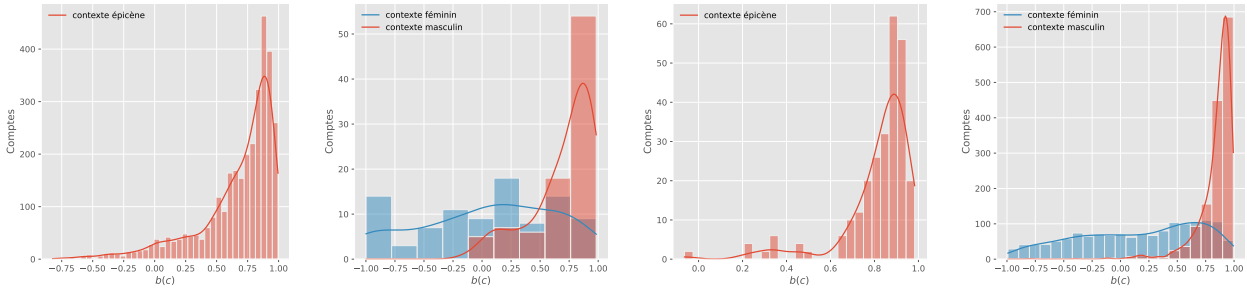
La figure 1 décrit la distribution de b lorsque les différentes probabilités sont estimées par un modèle de langue (donc en ignorant les informations provenant de la source) ; nous distinguons les cas où le nom de métier serait épicène en anglais des cas, plus rares, où le genre serait marqué. Dans les deux cas, le modèle a une tendance très claire à estimer pour `his` une probabilité très supérieure à celle de `her`. Pourtant, pour un modèle non biaisé, la valeur de b devrait être proche de 0 pour les phrases dans lesquelles aucune information de genre n'est exprimée (d'autant plus que, d'après l'hypothèse distributionnelle, `her` et `his` devraient avoir des représentations très proches). Elle devrait également montrer une préférence pour la « bonne » forme du pronom lorsque le genre est exprimé, alors que dans des contextes féminins b est en fait distribué sur l'ensemble de son domaine de définition, contrairement aux contextes masculins où les valeurs de b sont concentrées vers 1. Ces observations suggèrent que le modèle de langue soit n'est pas capable de capturer l'information de

3. Les résultats obtenus sur le motif (3)-(4) sont identiques, puisque ceux-ci ne repose que sur la comparaison des probabilités de générer les deux formes du pronom possessif.

4. Calculer b plutôt que le rapport entre les probabilités $p(\texttt{his}|c)$ et $p(\texttt{her}|c)$ nous permet d'assurer que toutes les valeurs sont comprises dans $[-1, 1]$ ce qui facilite les représentations graphiques.

genre dans le groupe sujet anglais soit n'est pas capable d'utiliser celle-ci pour choisir la forme du pronom possessif.

Si ces constatations ne sont pas complètement surprenantes au vu des nombres d'occurrences respectives de `his` et `her` dans le corpus d'apprentissage (cf. §3), elles montrent à quel point la prédiction de la bonne forme du pronom possessif par un modèle de traduction est compliquée : en plus de capturer correctement l'information dans la source et de « transférer » celle-ci, le modèle devra en plus « compenser » la tendance intrinsèque du contexte cible à prédire une forme masculine. Comme on peut le voir sur les figures 1c et 1d, la distribution de b lorsque la phrase source est considérée (c.-à-d. lorsque b est estimé par un modèle de traduction) a le même comportement que lorsque celle-ci est ignorée (c.-à-d. lorsque les probabilités sont estimées par un modèle de langue), montrant que la prise en compte de la phrase source ne permet pas de « corriger » le comportement du modèle de langue. Cette observation indique surtout que le système n'a pas besoin d'utiliser les mêmes informations pour générer `her` ou `his` : dans le cas de `her`, il est nécessaire que le système de traduction arrive à transférer l'information de genre depuis la phrase source, alors que la prise en compte du contexte cible est suffisante pour générer `his`.



(a) LM Genre non marqué — (b) LM Genre marqué — (c) TM Genre non marqué — (d) TM Genre marqué

FIGURE 1 – Biais de genre dans les modèles de langue de traduction tel que mesuré par b.

6 Impacts sur l'apprentissage

Les résultats présentés dans la section précédente indiquent que la prédiction de `his` et `her` ne s'appuie pas nécessairement sur les mêmes informations. Pour mettre en évidence l'impact de cette observation sur l'apprentissage du système de traduction, nous proposons de comparer l'impact des exemples féminins et des exemples masculins lors de l'apprentissage d'un système de traduction.

Plus précisément, nous avons considéré dans un premier temps toutes les phrases ayant un sujet féminin de notre corpus de test contrôlé, réalisé une étape d'apprentissage (passe *forward*, calcul de l'entropie croisée et passe *backward*) et avons extrait pour chaque *module* de l'encodeur et du décodeur le gradient accumulé lors de la rétro-propagation [5]. Nous avons ensuite répété cette manipulation en considérant cette fois uniquement les phrases ayant un sujet masculin [6]. Notons que

5. L'encodeur et le décodeur sont composés de différents *modules* `pyTorch`, il y a, par exemple, pour chaque couche de décodeur un module correspondant à l'auto-attention cible, un module correspondant à l'attention croisée avec la source, un module correspondant à une couche de *feed-forward* et un module correspondant à une couche de normalisation.

6. Nous avons ignoré 136 phrases de notre corpus ayant un groupe sujet épicène.

	couche	$\frac{\nabla \text{param}_{\text{masc}}}{\nabla \text{param}_{\text{fémi}}}$
décodeur	0	0.719517
	1	0.756060
	2	0.758951
	3	0.720758
	4	0.780754
	5	0.950173
encodeur	0	0.652739
	1	0.649395
	2	0.713145
	3	0.661217
	4	0.729006
	5	0.770339

(a)

where	layer	component	$\frac{\nabla \text{param}_{\text{masc}}}{\nabla \text{param}_{\text{fémi}}}$
decoder	0	dec_layer_norm	0.727845
		feed_forward	0.712666
		src_trg_att	0.714707
		trg_trg_att	0.752271
		x_layer_norm	0.684124
	5	dec_layer_norm	0.411318
		feed_forward	0.996350
		src_trg_att	0.630349
		trg_trg_att	0.708735
		x_layer_norm	0.638037
encoder	0	feed_forward	0.658266
		layer_norm	0.604996
		src_src_att	0.667691
	5	feed_forward	0.770884
		layer_norm	0.771792
		src_src_att	0.768297

(b)

TABLE 2 – Gradients accumulés après avoir vu les exemples masculins et féminins.

les phrases ayant un sujet féminin et les phrases ayant un sujet masculin ont sensiblement la même taille (respectivement 13 749 et 13 264 tokens).

La table 2 donne, pour les différentes couches de l'encodeur et du décodeur, le rapport entre les normes du gradient [7] obtenues à partir des phrases masculines et féminines. Nous avons également détaillé les valeurs de ce rapport sur les différents modules de la dernière couche de l'encodeur et du décodeur. Comme on pouvait s'y attendre, le gradient calculé sur les exemples féminins est systématiquement plus grand que celui calculé sur les exemples masculins (le système fait moins d'erreurs sur ces derniers), même s'il est intéressant de noter que la différence est plus marquée dans l'encodeur que pour le décodeur. Il apparait donc que, pour « corriger » les prédictions erronées des exemples féminins, le réseau de neurones cherche principalement à mieux extraire des informations de la source d'où peut être extraite l'information de genre.

7 Conclusion

Ce travail d'analyse des biais de genre participe, pour la traduction neuronale, du développement d'une culture numérique pour la traduction neuronale que Bowker & Ciro (2019) appellent de leurs voeux. Le travail met en évidence l'enjeu sociétal au prisme du biais de genre et s'inscrit dans les préoccupations majeures des humanités numériques. L'analyse des flux d'information participe d'une démarche plus générale qui contribue à l'intelligence artificielle explicable, bien qu'il existe d'autres approches du flux d'information Zuidema (2021).

7. Toutes les normes des gradients ont été normalisées par la taille du gradient (c.-à-d. le nombre d'éléments de celui-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 et par le projet SPECTRANS, dans le cadre de l'AAP émergence 2020 (ANR-18-IDEX-0001, Financement IdEx Université Paris Cité). Nous remercions le CNRS/TGIR HUMA-NUM et le Centre de Calcul IN2P3 pour la fourniture des ressources informatiques et de traitement des données nécessaires à ce travail.

Références

BARONI M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B : Biological Sciences*, **375**(1791), 20190307. DOI : 10.1098/rstb.2019.0307.

BOWKER L. & CIRO J. B. (2019). *Machine translation and global research : Towards improved machine translation literacy in the scholarly community*. Emerald Group Publishing.

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.

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.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 66–71, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KOPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J. & CHINTALA S. (2019). Pytorch : An imperative style, high-performance deep learning library. In H. WALLACH, H. LAROCHELLE, A. BEYGELZIMER, F. D'ALCHÉ-BUC, E. FOX & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 32*, p. 8024–8035. Curran Associates, Inc.

SAVOLDI B., GAIDO M., BENTIVOGLI L., NEGRI M. & TURCHI M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, **9**, 845–874. DOI : 10.1162/tacl_a_00401.

STANCZAK K. & AUGENSTEIN I. (2021). A survey on gender bias in natural language processing. *arXiv preprint arXiv :2112.14168*.

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.

WISNIEWSKI G., ZHU L., BALLIER N. & YVON F. (2021). Screening gender transfer in neural machine translation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, p. 311–321, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.blackboxnlp-1.24.

ZUIDEMA W. (2021). Language models, brains and interpretability. In *Plenary Talk for the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.