

Faculté	des Langues
---------	--------------------

Université de Strasbourg

**Master Technologies des Langues
Option Informatique-Traduction**

2019-2021

Défis posés par l'écriture inclusive au traitement automatique du français : le cas de la traduction automatique français-anglais

Sarah WEIGEL

Mémoire de deuxième année

Sous la direction de
Delphine BERNHARD

Responsable du Département d'informatique de la Faculté des Langues
Responsable pédagogique du Master 1 Technologies des Langues

Résumé : L'écriture inclusive (ou le langage inclusif) cherche à réintroduire le genre féminin dans un espace principalement masculin. Cette écriture passe par un certain nombre de règles et conventions qui modifient la langue au niveau morphologique et lexical. Ces changements linguistiques vont avoir des répercussions sur le traitement automatique du français à différents niveaux de la chaîne de traitement. Notre étude portant sur la traduction automatique de stratégies d'écriture inclusive en anglais par trois systèmes différents, nous révèle que globalement les systèmes parviennent à s'adapter à certaines formes d'écriture inclusive en utilisant diverses stratégies de traduction. L'inclusivité du texte est tout de même affectée, principalement lors de la traduction inversée où les formes inclusives sont peu maintenues.

Mots-clés : écriture inclusive, langage inclusif, traitement automatique des langues, traduction automatique

Table des matières

INTRODUCTION.....	1
1. L'ÉCRITURE INCLUSIVE, AMÉLIORATION ET MODERNISATION DE LA LANGUE.....	3
1.1. Les raisons d'un changement.....	3
1.2. Une morphologie et une syntaxe différentes.....	5
1.3. Des nouveautés morphologiques et une évolution lexicale.....	6
1.4. Vers une plus grande utilisation de la règle de proximité.....	8
1.5. Une évolution du neutre ou neutralisation.....	9
1.6. Synthèse des stratégies du langage inclusif en français et en anglais.....	11
2. LE TRAITEMENT AUTOMATIQUE DU FRANÇAIS ET SES CONFLITS AVEC L'ÉCRITURE INCLUSIVE.....	14
2.1. Un domaine vaste et encore en évolution.....	14
2.2. La tokenisation.....	15
2.3. La correction orthographique et grammaticale.....	16
2.4. La reconnaissance des mots inconnus.....	17
3. TRADUCTION AUTOMATIQUE DE L'ÉCRITURE INCLUSIVE.....	19
3.1. Les systèmes de traduction automatique.....	19
3.2. La place du genre dans la traduction automatique.....	20
3.3. La traduction automatique face aux stéréotypes.....	20
4. ÉTUDE DU COMPORTEMENT DES SYSTÈMES DE TRADUCTION AUTOMATIQUE NEURONAUX FACE À L'ÉCRITURE INCLUSIVE.....	23
4.1. Méthodologie.....	23
4.2. Analyse des formes d'écriture inclusive après traduction automatique.....	30
4.2.1. Utilisation des stratégies de traduction par les systèmes de traduction automatique.....	30
4.2.2. Analyse de l'écriture inclusive dans les traductions anglaises.....	44
4.2.3. Stratégies discursives après traduction inversée.....	46
4.3. Impact des systèmes de traduction automatique sur l'inclusivité.....	51
4.3.1. Présence du genre grâce aux pronoms.....	51
4.3.2. Progression quantitative de l'écriture inclusive.....	60
4.4. Conclusion de l'étude.....	64
CONCLUSION.....	66
RÉFÉRENCES BIBLIOGRAPHIQUES.....	67
ANNEXE.....	72
1. Abréviations utilisées.....	72

Table des figures

<i>Figure 1 : Capture d'écran de l'interface d'annotation des stratégies discursives en français de la plateforme Glozz.....</i>	<i>25</i>
<i>Figure 2 : Nombre d'occurrences des différentes stratégies discursives d'écriture inclusive</i>	<i>26</i>
<i>Figure 3 : Capture d'écran de l'interface d'annotation des stratégies de traduction de la plateforme Glozz.....</i>	<i>28</i>
<i>Figure 4 : Répartition de l'utilisation des stratégies de traduction pour le système de TA DeepL.....</i>	<i>34</i>
<i>Figure 5 : Répartition de l'utilisation des stratégies de traduction pour le système de TA Google Translate.....</i>	<i>38</i>
<i>Figure 6 : Répartition de l'utilisation des stratégies de traduction pour le système de TA Systran.....</i>	<i>42</i>
<i>Figure 7 : Répartition de l'utilisation des stratégies de traduction pour les trois systèmes de TA.....</i>	<i>43</i>
<i>Figure 8 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus français.....</i>	<i>52</i>
<i>Figure 9 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus traduit avec le système de TA DeepL.....</i>	<i>53</i>
<i>Figure 10 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus traduit avec le système de TA Google Translate.....</i>	<i>54</i>
<i>Figure 11 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus traduit avec le système de TA Systran.....</i>	<i>55</i>
<i>Figure 12 : Représentation du nombre de pronoms pour chaque catégorie dans le corpus d'origine et le corpus de traduction inversée avec le système de TA DeepL.....</i>	<i>56</i>

<i>Figure 13 : Représentation du nombre de pronoms pour chaque catégorie dans le corpus d'origine et le corpus de traduction inversée avec le système de TA Google Translate.....</i>	<i>57</i>
<i>Figure 14 : Représentation du nombre de pronoms pour chaque catégorie dans le corpus d'origine et le corpus de traduction inversée avec le système de TA Systran.....</i>	<i>58</i>
<i>Figure 15 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus aux différentes étapes de notre étude avec les trois systèmes de traduction automatique.....</i>	<i>59</i>
<i>Figure 16 : Matrice représentant la quantité de formes de chaque stratégie discursive par texte dans le corpus d'origine.....</i>	<i>63</i>
<i>Figure 17 : Matrice représentant la quantité de chaque stratégie discursive par texte après traduction inversée avec DeepL.....</i>	<i>63</i>
<i>Figure 18 : Matrice représentant la quantité de chaque stratégie discursive par texte après traduction inversée avec Google Translate.....</i>	<i>63</i>
<i>Figure 19 : Matrice représentant la quantité de chaque stratégie discursive par texte après traduction inversée avec Systran.....</i>	<i>63</i>

Table des tableaux

<i>Tableau 1 : Récapitulatif des stratégies du langage inclusif en français.....</i>	<i>12</i>
<i>Tableau 2 : Stratégies du langage inclusif en anglais.....</i>	<i>13</i>
<i>Tableau 3 : Nombre de tokens par article.....</i>	<i>24</i>
<i>Tableau 4 : Exemples d'utilisation des stratégies de traduction.....</i>	<i>29</i>
<i>Tableau 5 : Stratégie de traduction selon la stratégie discursive pour le système de TA DeepL.....</i>	<i>31</i>
<i>Tableau 6 : Stratégie de traduction selon la stratégie discursive pour le système de TA Google Translate.....</i>	<i>34</i>
<i>Tableau 7 : Stratégie de traduction selon la stratégie discursive pour le système de TA Systran.....</i>	<i>39</i>
<i>Tableau 8 : Nombre de formes d'écriture inclusive avant et après traduction selon le système de TA.....</i>	<i>61</i>
<i>Tableau 9 : Nombre de formes d'écriture inclusive avant et après traductions selon le système de TA et selon le texte.....</i>	<i>62</i>

INTRODUCTION

L'écriture inclusive (ou langage inclusif) est présentée dans sa forme la plus simple comme « l'ensemble des attentions graphiques et syntaxiques qui permettent d'assurer une égalité de représentations des deux sexes » (Sebagh & Baric, 2019, p. 4), ce qui montre une volonté de changement et de progrès. Cependant cette volonté n'a pas été très bien accueillie par tout le monde et l'écriture inclusive a été au cœur de nombreux débats ces dernières années avec des opinions très partagées. Par exemple, bien que le Haut Conseil à l'égalité entre les femmes et les hommes (HCE) ait recommandé en 2015 l'utilisation des formes abrégées avec le point final (une des stratégies du langage inclusif), en 2017 le Premier ministre et le ministre de l'Éducation en ont déconseillé l'usage. Plus récemment encore, la circulaire¹ publiée le 6 mai 2021 proscrit l'utilisation de l'écriture inclusive (en particulier du point médian) à l'école. Un des arguments souvent utilisé contre l'utilisation de cette écriture, est lié à la complexité qu'elle apporterait aux textes (écrits ou oraux). L'écriture inclusive serait alors seulement un frein langagier, et ne serait pas (pour certain·e·s) une forme véritable et acceptable d'évolution de la langue. Cette complexité, qu'elle soit syntaxique, orthographique ou morphologique, pourrait également avoir des répercussions dans d'autres domaines liés à la langue et au langage comme le traitement automatique des langues. Regroupant un grand nombre de procédés, le traitement automatique des langues (ou TAL) doit s'adapter à tous changements linguistiques pour ne pas donner des résultats obsolètes ou erronés.

Au centre des débats, l'écriture inclusive fut introduite de manière méthodologique et succincte par les éditions Mots-clés au sein d'un *Manuel d'écriture inclusives* sous la direction de Raphaël Haddad en 2016. Ce manuel, réédité et complété en 2019, a mis en marche de nombreuses réflexions sur le sujet (par exemple Charaudeau, 2018) . Cependant, il serait faux de penser que l'intérêt pour une féminisation de la langue n'existait pas avant 2016. On peut parler notamment du livre d'Eliane Viennot « Non le masculin ne l'emporte pas sur le féminin ! Petite histoire des résistances de la langue française » publié en 2014 mais également de l'article de Daniel Elmiger « Abréger les femmes pour mieux les nommer : féminisation de la langue et techniques abréviatives » dans la revue *Séméion : travaux de sémiologie* en 2008. Les problèmes que l'écriture inclusive pourrait poser étaient eux aussi déjà évoqués par (Gygax & Gesto, 2007). De manière générale, les questionnements sur le sexisme de la langue et de ses représentations était déjà traité depuis un moment avec un guide de féminisation dès 1999 (Becquer et al., 1999). Ainsi l'écriture inclusive, ou du moins la féminisation, est déjà depuis de nombreuses années un sujet important. Plus généralement, la langue et ses changements constants sont toujours source de réflexions (Abeillé et al., 2018; Elmiger,

1 <https://www.education.gouv.fr/bo/21/Hebdo18/MENB2114203C.htm>

2015). Ces travaux vont finalement dans la même direction que l'écriture inclusive, c'est-à-dire offrir des alternatives à la forte masculinisation de la langue française.

Si le TAL a fait l'objet de nombreux articles, on peut souligner la relative absence de travaux traitant du rapport direct entre les outils du TAL et l'écriture inclusive, ce qui montre que c'est un sujet nouveau encore assez peu développé.

En se focalisant cependant sur un outil particulier du TAL, la traduction automatique, on retrouve de nombreux travaux réalisés qui s'interrogent sur le genre et les stéréotypes. En 2013, déjà a lieu un questionnement sur le renforcement de stéréotypes par certains systèmes de TA populaires comme Google Translate (Ungerleider, 2013). En 2018, deux nouveaux articles (Prates et al., 2018; Vanmassenhove et al., 2018) approfondissent la réflexion sur la place du genre dans la traduction automatique et les effets sur les traductions obtenues. Ces travaux récents montrent bien que ce sujet est de plus en plus d'actualité dans divers domaines.

C'est ainsi que nous pouvons nous demander en quoi l'écriture inclusive peut-elle être un obstacle au traitement automatique du français et plus spécifiquement à la traduction automatique ?

Afin de répondre à cette question, nous allons tout d'abord nous intéresser à l'écriture inclusive à travers les nouvelles règles et conventions proposées ou réintroduites. Une fois que nous aurons mieux saisi l'écriture inclusive, nous examinerons comment elle peut entrer en conflit avec le vaste domaine du traitement automatique du français en présentant plusieurs outils. Afin d'aller plus loin dans le rapport entre le TAL et l'écriture inclusive, nous étudierons ensuite la traduction automatique qui est un outil déjà concerné par les questions de genre.

Enfin, nous réaliserons une étude sur le comportement de trois systèmes de traductions automatique (DeepL², Google Translate³, Systran⁴) face à un corpus d'articles de journaux et magazines en ligne annotés comportant diverses formes d'écriture inclusive. Cette étude nous permettra en particulier de voir les stratégies adoptées par les systèmes de traduction automatique pour traduire des formes d'écriture inclusive dans une langue non grammaticalement genrée (l'anglais), et quelles sont les conséquences de ces stratégies sur l'inclusivité dans les textes traduits.

2 Système de traduction automatique DeepL : <https://www.deepl.com/fr/translator>

3 Système de traduction automatique GoogleTranslate : <https://translate.google.fr/?hl=fr>

4 Système de traduction automatique Systran : <https://translate.systran.net/translationTools/text>

1. L'ÉCRITURE INCLUSIVE, AMÉLIORATION ET MODERNISATION DE LA LANGUE

1.1. Les raisons d'un changement

L'arrivée de l'écriture inclusive et d'une volonté de féminisation n'est pas anodine en français mais surtout elle n'est pas si récente. L'écriture inclusive n'est pas un débat nouveau comme le rappelle Julie Abbou dans l'épisode « Écriture inclusive : pourquoi tant de haine ? » du podcast de Binge Audio *Parler comme jamais*⁵. Depuis plus de 40 ans déjà on cherche à rendre la langue plus inclusive avec déjà le guide de féminisation publié en 1999 (Becquer et al., 1999) qui fait suite à des revendications féministes pour un accord en genre des métiers depuis les années 1970. Ce qui est nouveau c'est surtout le nom donné à cette initiative, « écriture inclusive » (ou plus largement « langage inclusif »), et certaines des stratégies proposées comme l'utilisation du point médian.

Si cette initiative de « démasculiniser » ou « réféminder » la langue a fait surface, ce n'est pas sans raison. La place du féminin et du masculin dans la langue peut s'observer de manière la plus évidente à travers les accords de genre. En français, tandis que le genre est immotivé pour les objets inanimés, il se distribue selon le sexe pour les noms désignant les animés. Cela explique le rapprochement évident du genre grammatical au sexe. Cependant, la prévalence du masculin sur le féminin découle de l'idée même que l'homme est supérieur à la femme (Houdebine-Gravaud, 2018, p. 219). C'est cette idée qui va régir les futures règles de grammaire depuis le XVII^{ème} siècle. La langue française n'a cessé de changer, d'évoluer avec de nombreuses réformes et l'introduction de règles parfois temporaires, mais malgré tout, la prévalence du masculin est demeurée, car elle n'est pas (ou peu) remise en question. Au contraire, on a pu voir la langue française subir une masculinisation importante découlant du sexisme de la société, entre le XVII^{ème} et le XIX^{ème} notamment sur les noms de métiers féminins qui existaient déjà à cette période.

Aujourd'hui, la langue n'appartient plus, pour certain·e·s, comme Florence Montreynaud (Mathot, 2018), aux carcans de l'Académie Française mais elle appartient aux locuteurs et locutrices, aux futures générations. L'écriture inclusive se veut donc la porte-parole d'un changement face à une tradition dépassée. L'écriture inclusive a pour but d'assurer une égalité de la représentation des deux sexes dans la langue. Ainsi, d'un point de vue plus concret, l'écriture inclusive cherche à renoncer au masculin générique et à la primauté du masculin sur le féminin dans les accords en genre. Cette écriture vise à diversifier la langue, et non à l'alourdir, afin de ramener le féminin dans un espace trop masculinisé (Sebagh & Baric, 2019, p. 4). Cette écriture est même une façon de revitaliser la langue

5 <https://www.youtube.com/watch?v=xj-Mn-jzOo0>

française. Il faudrait voir l'écriture inclusive comme un enrichissement de la langue mais surtout comme une modernisation. Il est important que les progrès réalisés dans nos sociétés soient également reflétés dans la langue, outil indispensable de la communication. Comme l'a dit Florence Montreynaud « La langue est le reflet de la société » (Mathot, 2018) ainsi la langue doit demeurer actuelle et moderne, car si les femmes et les hommes peuvent agir sur la langue, l'inverse est également vrai.

L'écriture inclusive va avoir un véritable effet sur les inégalités, car elle influence nos représentations. Il a été observé et montré, notamment par Gabriel *et al.* (2008), et par le sondage réalisé par Harris Interactive pour Mots-Clés en 2017 (Sebagh & Baric, 2019, p. 24), que la forme grammaticale d'un nom de métier pouvait influencer la représentation de ce métier. Il faut noter qu'à partir du moment même où on attribue un genre précis à une désinence ou à une unité syntaxique, on réactive les catégories de genre et de ce fait on entretient l'idée que le genre du référent est pertinent (Gabriel & Gygax, 2016, p. 1). On renforce ainsi de manière inconsciente l'existence de différences stéréotypées car le genre de la personne devient lié sémantiquement et lexicalement à la signification même du mot. Il a également été démontré que le masculin était en vérité peu perçu comme générique comme avec l'étude de (Gygax & Gabriel, 2012), montrant que l'activation de la valeur générique (ou neutre) du masculin passe par un traitement actif (réflexion consciente) de la forme en question tandis que la valeur spécifique du masculin, en tant que valeur par défaut, est activé par un traitement passif (réflexion inconsciente) de la forme. La féminisation peut influencer la représentation que se font les lecteurs et lectrices des protagonistes dans une phrase (Gygax & Gesto, 2007, p. 241). C'est grâce à ces nouvelles représentations que le combat pour l'égalité homme-femme peut avancer. L'écriture inclusive va notamment mettre au jour les « processus ordinaires d'invisibilisation des femmes » (Sebagh & Baric, 2019, p. 19). La réflexion imposée par l'utilisation et la mise en pratique de l'écriture inclusive dans nos rédactions nous permet de réaliser la présence ou plutôt l'absence des femmes au sein de la langue. C'est pourquoi Julie Abbou explique que c'est avant tout les débats que provoquent les réflexions sur l'écriture inclusive qui sont importants. Pour d'autres, c'est seulement à travers la normalisation de l'écriture inclusive et de son usage, dans les textes adressés à un grand public, que l'on pourra observer les plus gros changements (Tibblin, 2019, p. 47). Ce public est le premier concerné car comme le précise Julie Abbou (Binge Audio, 27/01/2021), la normalisation de l'écriture inclusive (ou du moins de certaines de ses stratégies) n'appartient pas aux linguistes mais bien de manière globale aux locuteurs et locutrices francophones.

Comme le rappelle Éliane Viennot (2018), « l'écriture n'est pas une finalité. Elle constitue un moyen(...) pour atteindre un objet civique, à savoir l'égalité » et de ce fait l'écriture inclusive dépasse le cadre même de la linguistique.

Pour certain·e·s comme Pascal Gygax (Binge Audio, 27/01/2021), l'importance de l'écriture inclusive serait même d'aller au-delà d'une « reféminisation » afin d'atteindre plus

d'égalité en se penchant vers une inclusivité non binaire et un véritable questionnement sur le genre dans la langue (française en particulier). Ce débat et ce questionnement qui est souvent nommé comme langage épïcène quand on s'intéresse plus à neutraliser ou retirer le genre dans la langue, met en avant des réalités considérées comme 'nouvelles' notamment la non-binarité.

Ainsi, à toute échelle l'utilisation de cette écriture a son importance et il est possible pour tout le monde de mettre en pratique les règles et conventions qu'elle propose.

1.2. Une morphologie et une syntaxe différentes

L'une des conventions majeures de l'écriture inclusive, et probablement celle qui a fait le plus de bruit, est celle liée aux changements graphiques et syntaxiques introduisant l'utilisation du point médian, ou bien des doubles formes. Ces changements visent à réintroduire le féminin (et de ce fait, la femme) au sein même d'expressions souvent masculines, ou vues comme neutres. Avec l'ajout d'une forme féminine directement d'un point de vue syntaxique, il devient impossible de ne pas prendre en compte la femme lors de l'interprétation du texte. Les deux genres grammaticaux doivent alors apparaître de manière égale, l'un à côté de l'autre dès que possible. Parmi les différents procédés permettant de former des noms communs de différents genres, on retrouve l'abréviation (Elmiger, 2008, p. 120) :

(1) *le/la prof, l'écolo, etc.*

L'abréviation garde généralement le début de la chaîne phonique ou graphique et entraîne la formation de mots épïcènes par la chute des terminaisons (Elmiger, 2008, p. 120). Les doubles formes abrégées représentent l'utilisation de ce procédé, et utilisent des signes marquant les frontières morphologiques, généralement soudées à la forme et non dissociées (Elmiger, 2008, p. 121) :

(2) *les étudiant·e·s, les directeurs/trices, les enseignant(e)s, etc.*

Il existe plusieurs signes et procédés typographiques utilisés pour former des doubles formes abrégées. Elmiger (2008) revient sur ces différents signes et procédés typographiques en soulignant les débats qu'ils suscitent. Parmi les plus utilisés, on retrouve les parenthèses qui sont, de manière générale, utilisées dans une phrase pour marquer les parties facultatives. Or, ce procédé est souvent critiqué par les partisan·e·s de l'écriture inclusive à cause de la connotation négative, plaçant les formes féminines entre parenthèses, leur donnant ainsi une valeur moins importante. La barre oblique est un autre signe typographique utilisé, qui permet généralement d'indiquer une équivalence entre les deux éléments séparés. Cependant, ce signe est aussi souvent rattaché à l'idée de séparation, voire d'exclusion, et il est ainsi souvent critiqué par les même partisan·e·s. Un autre signe typographique souvent utilisé est le trait d'union, car il souligne une cohésion des éléments, néanmoins, il peut aussi être utilisé comme quasi-parenthèse, ou afin d'introduire des répliques de dialogues. Ainsi, le point médian est aujourd'hui d'après de nombreux critiques, tels que Éliane Viennot, et de partisan·e·s de l'écriture inclusive, le signe typographique le plus adapté et celui qui est le plus recommandé. Il est préféré, car

il empêche la confusion avec le point final (un signe déjà très stabilisé dans l'usage) et ne connaît pas d'usage hérité (et encore en vigueur) ce qui lui permet de s'investir de l'enjeu discursif de l'écriture inclusive.

La composition et l'utilisation des doubles formes abrégées n'est pas aléatoire comme le présentent Sebagh *et al.* (2019). La structure à privilégier pour composer des doubles formes abrégées est 'mot masculin + point médian + suffixe féminin'. Cependant on doit utiliser cette méthode principalement lorsque les deux formes (masculin et féminin) se ressemblent c'est-à-dire lorsque l'ajout du suffixe féminin ne demandera pas un changement de la forme masculine. C'est le cas pour de nombreuses formes ayant un suffixe genré ou un suffixe féminin en -e se soudant naturellement à la forme masculine.

Les doubles formes abrégées ne sont cependant pas les seuls changements possibles parmi les conventions de l'écriture inclusive. Bien que les doubles formes abrégées soient préférées, leur utilisation n'est pas toujours possible lorsque la forme féminine diffère trop de la forme masculine. Ainsi, lorsque l'utilisation de ces formes est impossible, il est souvent conseillé, afin de ne plus seulement mentionner le masculin générique, d'avoir les formes féminines et masculines de manière juxtaposée (Elmiger, 2008, p. 121) :

(3) *les étudiants et étudiantes, ceux ou celles, les politiciens et politiciennes, etc.*

Ces deux formes, appelées doubles formes intégrales, apparaissent généralement par ordre alphabétique (Elmiger, 2008, p. 121). Elles permettent d'avoir la présence des deux genres, sous une seule idée, mais elles sont parfois vues comme redondantes.

1.3. Des nouveautés morphologiques et une évolution lexicale

Il existe d'autres conventions, liées au lexique qui peuvent permettre de varier les stratégies discursives de féminisation. Pour des raisons qui ne sont pas grammaticales, le féminin possède souvent une valeur péjorative lors de la création de nouveaux mots (Becquer *et al.*, 1999, p. 7). Ce n'est qu'avec l'usage et la normalisation que cette valeur pourra éventuellement changer. Néanmoins, la création de nouvelles formes permettant l'accord en genre a eu lieu et a déjà été encouragée, par exemple en 1999 par Becquer *et al.*. Pour la formation de ces nouveaux mots, on a une répartition des activités en trois catégories : (i) emplois manuels non valorisés, (ii) titres nobiliaires et charges, (iii) emplois valorisés. Selon la catégorie, on peut observer un processus de féminisation différent. Pour la féminisation des métiers de la première catégorie, on ne rencontre pas de problème particulier, sans doute car ce sont des métiers manuels, où les femmes secondaient leurs maris. Pour la deuxième catégorie, les titres et charges, que ce soit dans le domaine du droit ou de la religion, ont toujours été sexués, on a donc gardé les noms féminins aujourd'hui. Enfin, pour la dernière catégorie, on remarque une plus grande difficulté à la féminisation ce qui s'explique socialement avec l'idée que certaines activités sont 'réservées' à un certain genre. L'usage du déterminant ou l'emploi de formes épécènes va tout de même se banaliser progressivement.

Il est important de souligner que peu importe le domaine ou le type de nom que l'on crée, et bien que les formes épécènes se répandent de plus en plus, le but de l'écriture inclusive

et de la création de nouvelles formes est d'avoir (et de diffuser) des formes pleinement féminisées. Pour la création de noms féminins, plusieurs règles ont été mises en place. La règle qui concerne les noms se terminant au masculin par -e est que la forme au féminin sera identique (forme épïcène) mais on utilisera le déterminant féminin. Si le nom se termine par -é ou -i, on ajoute un -e à la finale pour obtenir la forme féminine. Bien que ces règles aient été mises en place pour homogénéiser et faciliter la création de formes féminines, il arrive que ces règles ne fonctionnent pas et on rencontre alors plusieurs problèmes. L'homonymie est l'un des premiers obstacles, car certains noms de métiers féminisés existent déjà pour désigner autre chose :

(4) *La crêpière, la dépanneuse, la jardinière, la vernisseuse, etc.*

Cependant, en contexte, la polysémie ne pose pas de problème. Un autre obstacle, plus sociologique, est la dévalorisation, car pour beaucoup de femmes le nom du métier au féminin revient à dévaloriser le métier. Cela peut être lié à l'euphonie (impression de dissonance d'un mot) car souvent certains hauts postes étaient réservés aux hommes et ainsi la valeur même du poste fut associée au nom du métier au masculin. Il faut souligner le fait que bien que la notion de beauté est utilisée par les opposant·e·s de l'écriture inclusive, cela est purement subjectif et que c'est l'usage et l'habitude qui pourront influencer cette idée (Mathot, 2018). Néanmoins, un autre problème à soulever qui est souvent lié à cette dévalorisation est le fait que souvent les noms féminins des homologues masculins ont des connotations péjoratives et/ou sexuelles telles que « maîtresse », « coureuse » ou « entraîneuse » (Viennot et al., 2016, p. 84). Ce problème qui dépasse quelque peu le domaine linguistique rappelle le problème de la dépréciation du féminin et sa banalisation.

Enfin bien que l'accord des noms de fonctions, grades, métiers et titres en genre est possible, les formes féminines sont souvent encore peu usitées. Ces formes sont souvent présentes et parfois même depuis longtemps comme par exemple la forme *autrice* qui fait parfois débat. Ce nom féminin pour qualifier « une femme qui a écrit un ou plusieurs ouvrages, de quelque nature que ce soit » (linternaute.fr) n'est pas aussi récent que l'on imagine. Evain (2008) retrace l'histoire de ce nom et son parcours dans la langue. On retrouve les premières occurrences d'« autrice » (sous la forme *auctrix*) déjà au cours des premiers siècles du christianisme et cet usage continuera notamment pour décrire l'origine d'un mouvement ou d'une action. Tandis que certains affirment que « auteur » (sous la forme *auctor*) peut être utilisé pour les deux sexes, l'usage de la forme féminine ne disparaît pas pour autant avec les premières occurrences du terme en littérature aux X^{ème} et XII^{ème} siècles. C'est seulement au XVI^{ème} siècle que l'on retrouvera la première occurrence du terme ayant le sens d'écrivaine ; or le siècle suivant on tente peu à peu de faire disparaître ce terme qui renvoie à une réalité qui dérange, celle des femmes de lettres dans un monde où la professionnalisation du champ littéraire a lieu. C'est ainsi qu'au XVIII^{ème}, le terme « autrice » est supprimé des manuels de langue sous l'autorité de l'Académie. Le terme revient en force avec les nouvelles questions sur la place de la femme dans la langue. A la fin du XIX^{ème}, pour la première fois « féminisation et

légitimisation vont de pair » (Evain, 2008) ce qui permet de faire avancer la question de l'usage ce terme. C'est seulement en 1996, que le terme *autrice* est de retour dans les dictionnaires (dans *le Petit Robert*). L'histoire de ce terme permet de rappeler que la création et l'usage de formes féminines n'est pas un 'phénomène' récent mais bien une réalité langagière.

Dans d'autres domaines, la création de néologismes est aussi possible à l'aide d'un suffixe épïcène (forme invariable) ce qui permet d'obtenir une forme neutralisée ou du moins non genrée, comme le montrent exemples repris par Tibblin (2019) :

(5) *Professoraire ou professoriste, écrivainaire ou écrivainiste, etc.*

Il est également possible d'avoir recours aux termes épïcènes ou aux formulations englobantes afin de renvoyer au masculin et au féminin, comme le présentent les exemples du *Manuel d'écriture inclusive* (2019) :

(6) *L'artiste, l'interprète, les propriétaires, la direction etc.*

On utilise alors une forme plus neutre qui va pouvoir renvoyer aux deux genres indépendamment. Ces créations font partie des différentes conventions et stratégies que préconisent les partisan·e·s de l'écriture inclusive. Or pour rendre la langue française plus inclusive, il existe déjà des moyens que l'on a décidé d'oublier et que l'on pourrait remettre au goût du jour.

1.4. Vers une plus grande utilisation de la règle de proximité

L'une des stratégies du langage inclusif, qui ne demande pas de changements syntaxiques ou morphologiques, est la plus grande utilisation de la règle de proximité. Les stratégies d'accord en français au sein du groupe nominal sont l'accord en nombre et l'accord en genre. La règle de résolution est la règle d'accord qui est utilisée de manière générale et consiste à accorder un ou plusieurs noms selon le genre intrinsèque donné. Néanmoins, lors de l'accord entre deux noms de genres différents, un choix doit être fait entre le masculin et le féminin et aujourd'hui l'accord au masculin reste la norme avec l'idée que le « masculin est supérieur au féminin ». Pour l'accord en genre avec les adjectifs par exemple, on a selon Corbett (Abeillé et al., 2018, p. 5), une résolution au profit du masculin peu importe le cas.

Cependant, cela n'a pas toujours été le cas, et de nombreuses associations (telles que L'égalité c'est pas sorcier, Le monde selon les femmes), ainsi que la maison d'édition indépendante Editions iXe, essaient de relancer la règle dite « règle de proximité » qui fut utilisée du XV^{ème} au XIX^{ème} siècle. La règle de proximité consiste à avoir une approche différente de la coordination de genres différents. On va s'intéresser à la syntaxe et ainsi réaliser un accord avec le genre du nom le plus proche. L'accord de proximité dans le français contemporain existe bien, car après deux articles de genres différents, on remarque que l'accord du nom se fera avec l'article le plus proche (Abeillé et al., 2018, p. 8) :

(7) [...] *j'appelle un ou une amie* [...]

(Frantext : Pierre Pachet, *Devant ma mère*, Paris, Gallimard, 2007, p. 136)

(8) [...] *ceux qui ont deux grand-parents juifs et ont épousé un ou une juive* [...]

(Frantext : Pierre Mendès-France, *S'engager, 1922-1943*, in *Œuvres complètes*, Paris, Gallimard, t. I, 1984, p. 558)

Idem lorsqu'on a deux noms de genres différents, l'article sera épïcène ou bien s'accordera avec le genre du nom le plus proche (Abeillé et al., 2018, p. 15) :

(9) *On voit en leur lieu et place des corps mutilés* [...]

(Frantext : Marguerite Duras, *Hiroshima mon amour* [1960], Paris, Gallimard, 1994, p. 9)

(10) *Présentez-vous à la date et lieu indiqué pour suivre votre formation.*

(FrWaC : secours57.fr)

Cette règle est donc déjà en vigueur et également légitime, car selon Curat (Abeillé et al., 2018, p. 5), l'accord de résolution au masculin n'est pas toujours acceptable et parfois l'accord de proximité peut être préféré.

Cette question d'acceptabilité pose cependant souvent problème, car bien que grammaticalement l'accord fonctionne, les normes et habitudes discursives entrent en conflit avec l'accord réalisé. C'est pourquoi, souvent, on préconise de mettre la règle de résolution et la règle de proximité sur le même plan en plaçant le mot masculin au plus près de l'adjectif afin que les deux règles coïncident. Or, en réalisant cela, on continue de masculiniser la langue en tentant de manipuler une règle qui pourrait permettre un changement. C'est l'ordre des mots qui devient alors important dans la phrase et qui va même permettre de transmettre un certain point de vue. En effet, lors de l'utilisation de déterminants féminin et masculin à la suite, on retrouve plus fréquemment l'ordre « masculin-féminin ». Cependant, si l'on ne change pas l'ordre des mots pour avoir une correspondance entre les deux règles, lorsqu'il y a un conflit, il est recommandé de privilégier l'accord de proximité. Or, cela ne fonctionnera pas pour l'ordre des noms humains où le genre masculin continue de l'emporter. Si on rencontre des problèmes de coordinations de noms de genres différents, il est possible d'utiliser le pluriel ou des articles épïcènes.

1.5. Une évolution du neutre ou neutralisation

On observe ainsi la possibilité d'une adaptation du lexique notamment grâce au neutre et à la neutralisation. Il est d'abord important de différencier le genre grammatical (caractéristique intrinsèque des noms) et le genre lexical (propriété sémantique du nom désignant un référent animé sexué). En français, le genre grammatical et le genre lexical coïncident pour les référents animés sexués (Elmiger, 2015, p. 2). Pour les noms communs inanimés, la relation entre le nom et son genre grammatical est plus compliquée. On peut observer des régularités formelles ou des régularités dans certains champs lexicaux, mais globalement le genre grammatical n'est pas motivé. Pour les noms désignant des référents animés cependant, il existe souvent deux formes (distinctes ou proches), ou une seule forme utilisable avec les deux genres.

On parle de langue grammaticalement genrée lorsque les formes des mots reliés aux noms sont influencées par le genre même du nom ce qui est bien le cas en français (Gabriel & Gygax, 2016, p. 4). L'anglais, au contraire, se situe dans un entre deux car bien

que l'on retrouve le genre par les pronoms personnels, le genre n'est pas grammaticalement marqué (Gabriel & Gygax, 2016, p. 4). L'anglais est alors une langue genrée au niveau pronominal. Aujourd'hui 112 des 257 langues incluses dans le World Atlas of Language Structures (WALS) ont un système genré, et pour 84 de ces langues, ce système est basé sur le sexe du référent (Gabriel & Gygax, 2016, p. 4). Tandis qu'on retrouve des marques du neutre (comme le « o ») dans d'autres langues romanes, cela n'est pas le cas en français ce qui remet en question l'existence d'un genre neutre. Le neutre en français est en vérité le résultat d'un processus de neutralisation englobant les deux catégories du genre. Afin de démontrer ce processus de neutralisation, on peut prendre plusieurs exemples dans la langue (Charaudeau, 2018). Tout d'abord en ce qui concerne les déterminants et les articles, l'article « le » provient du latin *ille* (forme masculine singulière) et *illum* (forme neutre). Il peut ainsi désigner à la fois le genre masculin et le neutre. Ce procédé est identique pour l'article « un » qui possède aussi cette double identification. Pour les déterminants, la forme « ce » peut aussi désigner le genre masculin ou le neutre. Il existe donc des formes neutres, mais on a plutôt affaire à des cas de polysémie. Un autre exemple concerne les pronoms personnels, particulièrement les pronoms « il/on ». Le pronom « il » peut être parfois impersonnel, ne reprenant aucune personne, ce qui est une forme de neutralisation. Le pronom « on » peut aussi posséder un rôle de neutralisation en renvoyant à diverses entités. Une forme de neutralisation existe donc et permet l'indifférenciation entre les genres.

On peut observer différentes stratégies particulières de neutralisation. D'abord de manière générale, le processus de neutralisation discursive consiste « à faire perdre aux mots une partie de leur charge sémantique » (Charaudeau, 2018, p. 22). En général, la forme employée durant ce processus sera celle du masculin et permettra de désigner un ensemble. Cette forme de généralité va permettre de créer des concepts, où la forme masculine désigne une notion abstraite, une généralité englobante. Un exemple de cette neutralisation discursive est le mot « homme » qui fait encore souvent débat. Dans son origine latine, le mot permet de désigner l'être humain. Or, c'est un cas de polysémie, car ce mot désigne aussi l'être masculin. Son utilisation englobante, pour signifier les êtres humains au singulier ou au pluriel, permet au mot de représenter une catégorie conceptuelle non genrée. En sémantique, la neutralisation correspond aux « cas d'opposition où au moins l'un des deux éléments peut s'utiliser avec ou sans un trait différentiel, pertinent au niveau du cas particulier, mais non opératoire à celui de la catégorie plus générale. » (Elmiger, 2015, p. 6). On retrouve sous cette définition-là l'utilisation du masculin générique. Le masculin générique correspond à l'utilisation de la forme masculine à laquelle on attribue une valeur générique. Cette règle est encore aujourd'hui prescrite dans les grammaires normatives. On considère ainsi le masculin comme « un genre non marqué » (Elmiger, 2015, p. 5) or cela est souvent remis en question, car bien que cette forme se veuille neutre, elle garde tout de même une apparence masculine.

Le processus de neutralisation formelle concerne lui les « paires de noms apparentés d'un point de vue morphologique » (Elmiger, 2015, p. 7) où on peut retrouver une forme variant entre masculin et féminin selon le contexte. Ces formes sont généralement des formes épïcènes c'est-à-dire des formes qui n'ont pas de genres définis intrinsèquement mais vont prendre un genre selon le contexte. Par exemple, on peut retrouver des noms de métiers qui sont épïcènes et qui doivent leur genre aux articles.

(11) *Le journaliste/la journaliste, le peintre/la peintre, etc.*

Ces formes épïcènes sont vues comme beaucoup comme la seule véritable preuve et possibilité de neutralisation car elles permettent de désigner toutes personnes même en dehors des limites binaires (hommes/femmes).

Bien que la neutralisation soit une stratégie discursive possible pour une écriture plus inclusive, elle ne fonctionne pas dans tous les cas. Pour les noms communs inanimés, le genre est aléatoire, on ne peut donc pas parler de neutralisation tandis que pour les noms communs désignant une personne singulière, le générique (permettant une neutralisation) ne peut être utilisé, car on nierait la singularité de la personne. Il existe tout de même des propositions de nouvelles formes de neutralisations pour contrer certains problèmes. On retrouve d'abord une neutralisation formelle préconisant la création de formes épïcènes pour les noms, adjectifs et formes verbales homophones à l'oral.

(12) *Les professionnelles, venez, avertiz*

(Pour une grammaire non sexiste, Céline Labrosse 1996, p.95)

Dans la même approche, on retrouve la création de pronoms uniques (au singulier ou au pluriel) qui pourraient être utilisés à l'écrit ou à l'oral (Elmiger, 2015, p. 8-9). Ces formes seraient le résultat d'une fusion de deux formes. Ainsi la Suède (le suédois est une langue où le genre grammatical n'existe pas) adopte progressivement le nouveau pronom, « hen », afin d'avoir une alternative neutre (Darbois, 2017). Cela fait écho à l'idée, que soutient Pascal Gyax, qu'il ne faudrait pas démasculiniser la langue mais bien la dégenrer notamment car un grand nombre de personnes ne se retrouvent pas dans la binarité homme/femme. Ainsi en français, le pronom « iel » fait peu à peu son apparition et pourrait être une alternative possible aux pronoms binaire « il » et « elle ».

Il est parfois difficile de ne pas fournir d'informations concernant le genre du référent même quand cela n'est pas pertinent. Pour autant, il est possible de ne pas seulement référencer le genre masculin (Gabriel & Gyax, 2016, p. 5).

1.6. Synthèse des stratégies du langage inclusif en français et en anglais

A l'aide d'un Tableau 1, nous pouvons voir les principales stratégies du langage inclusif en français. Ce tableau reprend les différentes stratégies que nous avons présentées auparavant, et permet à l'aide d'exemples de mieux comprendre leur fonctionnement. Il met aussi en avant le fait que différentes solutions et alternatives sont possibles, afin de s'adapter à toutes les situations langagières.

Stratégies du langage inclusif (français)	Exemples
Les doubles formes intégrales	<i>Les français et françaises, elles et ils</i>
Les doubles formes abrégées	<i>l'intellectuel·le, un·e lycéen·ne</i> (Sebagh et al., 2019)
Les créations langagières : néologisme et féminisation	<i>lel, ceux, l'acaémicienne, l'écrivaine</i> (Tibblin, 2019: 3 ; Becquer et al., 199)
Les épïcènes et formulations englobantes	<i>Les juristes, les humains, le/la capitaine</i> (Sebagh, et al, 2019)
La règle de proximité	<i>A l'heure où la France s'apprête à élire un ou une nouvelle présidente [...]</i> (FRWAc : zideesdemars.com, 27 février 2012)

Tableau 1 : Récapitulatif des stratégies du langage inclusif en français

Comme nous allons utiliser l'anglais dans notre étude, nous nous sommes également intéressée au langage inclusif en anglais et aux stratégies disponibles. Parmi ces stratégies, utilisées notamment par les Nations Unies dans la rédaction de documents en anglais⁶, on retrouve les doubles formes intégrales, principalement pour les pronoms mais également pour les noms communs quand il y a plusieurs formes disponibles qui dépendent du genre. L'utilisation du pluriel et des formulations englobantes est également encouragée avec le pronom personnel « they » utilisé au singulier pour une désignation neutre ou non genrée, ainsi qu'avec les alternatives non marquées (utiliser « police officer » au lieu de « policeman »⁷). De plus, quand un genre en particulier n'est pas pertinent, il est possible de remplacer la forme pronominale par un nom ou par une forme renvoyant à la seconde personne du singulier, voire même de supprimer la forme en utilisant plutôt une formulation impérative. Une des autres stratégies encouragées est l'usage d'un langage non discriminant, c'est à dire qui ne renforce pas des stéréotypes ou des biais (éviter les expressions comme « In a manly way », « She runs like a girl ») et au contraire utiliser des formes plus neutres (épïcènes).

6 <https://www.un.org/en/gender-inclusive-language/guidelines.shtml>

7 <https://writingcooperative.com/how-to-make-your-writing-more-gender-inclusive-1961686124f3>

Le Tableau 2 ci-dessous reprend quelques stratégies disponibles en anglais pour assurer un langage plus inclusif.

Stratégies du langage inclusif (anglais)	Exemples
Doubles formes intégrales	<i>He or she, boys and girls</i>
Pluriel et formulations englobantes	<i>They, humankind, police officer</i>
Langage non discriminant et épiciènes	<i>Guests are cordially invited to attend with their partners.</i>

Tableau 2 : Stratégies du langage inclusif en anglais

Les changements que nous avons présentés à travers les stratégies discursives d'écriture inclusive sont vus par beaucoup comme des changements forcés et non conséquence d'une véritable évolution naturelle. De fait, ces changements pourraient poser de nombreux problèmes en linguistique et en informatique, car on fait face à une intervention directe dans la langue qui va demander une réaction (adaptation ou refus).

2. LE TRAITEMENT AUTOMATIQUE DU FRANÇAIS ET SES CONFLITS AVEC L'ÉCRITURE INCLUSIVE

Les bouleversements qu'implique l'écriture inclusive en linguistique vont avoir des répercussions sur d'autres domaines tel que le traitement automatique des langues (TAL). Les obstacles que peut poser l'écriture inclusive sont visibles à plusieurs niveaux du traitement automatique de la langue et peuvent empêcher l'obtention de résultats pertinents. Néanmoins, le TAL étant un domaine en constante évolution, des solutions sont souvent déjà disponibles afin de s'adapter à ces nouvelles formes.

2.1. Un domaine vaste et encore en évolution

Le TAL est un domaine vaste et encore en évolution notamment face aux nombreux changements de la langue comme pourrait l'être la féminisation qu'implique l'écriture inclusive. Le TAL, selon la définition qu'en donnent Bouillon et al. (1998), « a pour objet la création de programmes informatiques capables de traiter automatiquement les langues naturelles. ». Il faut compléter cette définition aujourd'hui, car le TAL passe par la conception de véritables modèles informatiques et surtout a aussi pour but d'exploiter les données récupérées. La conception d'un système de TAL suit un enchaînement de trois modèles principaux, selon Lebarbé (2015) : un modèle linguistique (issu d'une théorie expliquant le fonctionnement de la langue), un modèle informatique (théorie informatique utilisée afin de mettre en application le modèle linguistique) et un modèle de représentation (représenter les résultats obtenus par le modèle informatique). L'écriture inclusive risque de poser problème dès l'élaboration du modèle linguistique, car bien que des conventions aient été fixées, on remarque encore une instabilité dans l'usage et une forme d'hésitation (notamment vis-à-vis des signes typographiques pour les doubles formes abrégés). Ce manque de normalisation risque de rendre l'élaboration d'une théorie commune expliquant ces nouvelles formes compliquées. Au niveau du modèle informatique et du modèle de représentation, à nouveau, l'écriture inclusive demande un changement important, car elle apporte des changements considérables à la syntaxe usuelle de la langue française, et demande aussi une réinterprétation de certains mots (comme les mots épïcènes).

Les conflits entre le TAL et l'écriture inclusive sont donc inévitables, cependant ces conflits ne doivent pas être nécessairement perçus de manière négative. C'est par la construction d'un système de TAL que l'on a accès à de « nouvelles connaissances de la langue par l'analyse des imperfections mêmes du système » (Lebarbé, 2015, p. 466). Les problèmes rencontrés seront sources d'une réflexion plus importante sur les connaissances linguistiques, ce qui va permettre de refaçonner l'outil de TAL en question. Les nouveaux contextes de formes imparfaites représentent aussi des « nouveaux territoires à explorer » (Lebarbé, 2015, p. 473) pour le TAL et doivent être vus comme une source de nouvelles réflexions sur les modèles existants. Cette réflexion est importante, car on n'a pas encore

affaire à des résultats complètement fiables. En effet, les modèles sont encore imparfaits même face à des formes dites « correctes » (conforme à la grammaire prescriptive) d'énoncés.

2.2. La tokenisation

L'un des outils du TAL où l'écriture inclusive peut poser problème est l'outil de tokenisation. On parle de tokenisation pour décrire le processus visant à identifier et séparer les unités d'un texte. La tokenisation permet d'obtenir deux types de tokens : les unités où la structure du caractère est reconnaissable (ponctuations, nombres, dates) et les unités qui doivent passer par une analyse morphologique (Grefenstette & Tapanainen, 1994, p. 1). On ne peut pas parler de mots, car cette notion est différente entre la conception intellectuelle du linguiste et la conception numérique d'une machine. Lors de la tokenisation, la relation entre les unités n'est pas étudiée en tant que telle, mais c'est plutôt la séparation de ces éléments qui est mise en avant. La tokenisation est généralement la première étape lors du traitement automatique d'une langue, car elle permet aussi d'identifier les séparations de phrase. Cette division est importante, car généralement, la phrase est la base sur laquelle certaines analyses seront réalisées plus tard. De ce fait, il faut souligner que les choix qu'implique la tokenisation vont avoir des répercussions sur toutes les futures analyses de l'unité.

Afin de reconnaître une phrase, on peut se baser sur la ponctuation (Grefenstette & Tapanainen, 1994, p. 4). Cependant, tandis que les points d'exclamation et d'interrogation sont rarement ambigus, le point pose souvent problème. Le point peut représenter un point final, mais il peut aussi faire partie d'une abréviation, voire les deux. Certains tokens contiennent dans leur structure même un point comme les nombres, les références alphanumériques, les dates, les acronymes ou les abréviations. Les outils de tokenisation vont être capable de reconnaître certains de ces cas. En ce qui concerne les nombres, les expressions régulières arrivent généralement à identifier les constructions numériques particulières (Grefenstette & Tapanainen, 1994, p. 4). Les abréviations, qui sont souvent instables, posent plus de problèmes. Néanmoins, les expressions régulières, se basant sur différents cas possibles d'abréviations, peuvent aider à diminuer les erreurs afin de reconnaître un grand nombre d'unités.

Cette reconnaissance par les outils de tokenisation demande néanmoins une analyse pointue et constante des différentes formes possibles. Les expressions régulières doivent être mises à jour et l'arrivée de l'écriture inclusive demande non seulement d'importants changements, mais une toute nouvelle vision des signes typographiques tels que le point, le tiret ou les parenthèses. De plus, le point médian, qui est préconisé, n'est pas encore dans l'usage ainsi les outils de tokenisation vont faire face à un signe inconnu et des cas impossibles à traiter correctement. Un autre problème, qui ne touche pas la ponctuation, est celui de la correspondance d'une unité à une classe. Il arrive qu'un token soit en réalité composé de deux éléments syntaxiquement différents et cela peut poser problème par la suite notamment lors d'analyses morphologique et grammaticale. Il existe des arguments

allant dans les deux sens, c'est-à-dire soit séparer l'élément en plusieurs tokens, ou le considérer comme un seul token (Grefenstette & Tapanainen, 1994, p. 8). On peut rattacher ce problème aux doubles formes abrégées qui, si traitées comme un seul token, peuvent poser problème par la suite. Il est ainsi important de décider dès la tokenisation comment on décide de prendre en compte le suffixe féminin. Chaque système peut rendre un résultat différent et cette tokenisation influencera la suite des analyses du texte comme avec l'analyse morphosyntaxique. Cette analyse permet d'évaluer les tokens reconnus (avec leurs propriétés morphologiques notamment) mais également les propositions syntaxiques. Ainsi si la tokenisation comporte des erreurs, cette analyse ne pourra pas avoir lieu ou du moins pas correctement.

2.3. La correction orthographique et grammaticale

Un autre outil du TAL qui peut entrer en conflit avec l'écriture inclusive est celui de la correction orthographique et grammaticale. Un correcteur orthographique est destiné à détecter des fautes d'orthographe (Fontenelle, 2005, p. 120) : ainsi, il ne demande pas d'informations concernant les propriétés syntaxiques des mots contrairement à un correcteur grammatical. Un correcteur orthographique nécessite un dictionnaire dont l'élaboration peut poser problème notamment à cause de la morphologie dérivationnelle (Fontenelle, 2005, p. 120). Toutes les différentes formes doivent être présentes dans le dictionnaire pour être reconnues. De plus, il faut tenir compte du lexique attesté ou non, et ainsi choisir d'inclure ou non des mots qui sont courants dans l'usage, mais peut-être pas encore officiellement reconnus. L'écriture inclusive et les nombreuses créations de formes féminines (ou épïcènes) font donc partie de ce lexique qui n'est pas toujours, ou du moins pas entièrement, attesté. Pour choisir quels termes feront partie du dictionnaire, on s'intéresse à la notion de fréquence (Fontenelle, 2005, p. 120). Plus la fréquence d'utilisation d'un mot sera haute, plus le mot se retrouvera dans le dictionnaire du correcteur.

Les noms propres posent souvent problème, car un correcteur ne va pas pouvoir se baser sur la seule présence de la majuscule, et nécessitera alors une grande liste de noms propres d'entités nommées. Ainsi, les changements dans les normes orthographiques influencent également les programmes de correction orthographique, notamment avec l'existence de deux graphies qui peuvent encore être acceptées. En effet, cela demandera l'introduction d'informations supplémentaires dans les dictionnaires si on veut différencier l'orthographe traditionnelle de la nouvelle orthographe (Fontenelle, 2005, p. 122). Les changements qu'apporte l'écriture inclusive vont alors entrer en conflit avec les données déjà présentes dans les dictionnaires qui ne contiennent pas encore les différentes formes utilisées. Les doubles formes abrégées ne seront pas reconnues comme correctes, de par leur nature composite, et les nouveaux mots ne seront pas encore acceptés et seront ainsi jugés inconnus.

Dans ce cas-là, ce problème influencera l'autre rôle du correcteur orthographique qui est de fournir la ou les suggestions les plus probables pour corriger les fautes (Fontenelle,

2005, p. 123). Différentes opérations peuvent être la source d'une faute (suppression, ajout, permutation, remplacement) et le correcteur orthographique doit pouvoir proposer la ou les solutions appropriées selon le cas. Ce sont les données entrées dans le correcteur qui en font tout son fonctionnement et pour un correcteur grammatical, les informations sont encore plus nombreuses et dépassent les informations morphosyntaxiques de base telles que le genre et le nombre. On retrouve par exemple, les indications relatives aux propriétés de transitivité des verbes, aux prépositions liées à un verbe, aux complétives possibles (Fontenelle, 2005, p. 125). On retrouve aussi des indications sémantiques qui permettent d'identifier les noms de métiers ou les parties du corps. L'ajout de nouveaux noms de métiers féminins sera facilité, mais demandera tout de même une importante réflexion sur les formes possibles et utilisées qui ne sont encore pas toujours fixes. Ainsi, le dictionnaire utilisé par le correcteur (orthographique ou grammatical) est l'élément le plus important de l'outil. L'écriture inclusive en introduisant ses propres règles (orthographiques, syntaxiques) va alors entrer en conflit avec les précédentes connaissances, et l'identification des fautes sera incorrecte ou impossible. Des progrès ont cependant été réalisés autour de la question des mots inconnus.

2.4. La reconnaissance des mots inconnus

On parle de « mots inconnus » ou formes « hors vocabulaire » pour qualifier les mots absents d'un lexique utilisé pour le TAL (Spriet *et al.*, 1996, p. 1). La rencontre de mots inconnus lors du traitement d'un texte est inévitable même si on met l'écriture inclusive de côté. Les noms propres, les néologismes, les mots composés et les fautes de frappe sont déjà des exemples de formes posant problème aux outils du TAL. Spriet *et al.* (1996) décomposent les formes hors vocabulaire en trois classes (composés, noms propres, mots classiques) car elles demandent chacune un traitement spécifique. Du côté des méthodes, il en existe plusieurs qui peuvent permettre de remédier aux erreurs. On retrouve la méthode à base de règles (syntaxiques ou morphologiques), la méthode probabiliste (modèle triclassées ou trilettes) et la méthode à base d'arbre de décision.

L'approche trilette de la méthode probabiliste peut s'appliquer aux formes de la classe des mots classiques où l'on retrouve principalement des néologismes. Cette approche se base sur l'enchaînement de trois lettres (principalement en fin de mots) et les probabilités d'apparition de cet enchaînement selon les catégories grammaticales (Spriet *et al.*, 1996, p. 4-5). Cette approche permettra grâce aux probabilités récupérées d'attribuer certaines étiquettes aux mots inconnus. La méthode avec arbre de décision permet d'observer la position des lettres et la construction d'une terminaison des mots à base d'une liste de questions simples (Spriet *et al.*, 1996, p. 7-8). À nouveau, cette approche permet de définir les étiquettes appropriées pour un mot inconnu. Ces méthodes montrent donc qu'il existe déjà des possibilités de traitements de mots inconnus comme le seraient certains mots introduits par les conventions et créations du langage inclusif.

L'écriture inclusive peut bien poser des problèmes au TAL et est parfois même un obstacle à la bonne utilisation de certains outils. Cependant l'écriture inclusive permet d'abord une

réflexion sur la place du genre (notamment le genre féminin) dans le TAL ce qui peut être très intéressant pour certains outils comme la traduction automatique où les représentations ont une place importante.

3. TRADUCTION AUTOMATIQUE DE L'ÉCRITURE INCLUSIVE

L'importance du genre et des représentations se retrouvent déjà dans les réflexions récentes sur la traduction automatique (TA). L'écriture inclusive va permettre d'approfondir ces réflexions mais aussi de soulever les éventuels problèmes qui demeurent encore.

3.1. Les systèmes de traduction automatique

La TA telle que nous la connaissons aujourd'hui fut pendant longtemps un rêve. C'est à partir du XX^{ème} siècle que ce rêve put devenir réalité avec les progrès réalisés dans les divers domaines où la traduction automatique tire ses origines : la linguistique et l'informatique puis aussi l'intelligence artificielle (Hutchins & Somers, 1992, p. 3).

Aujourd'hui, bien que les résultats obtenus ne sont pas toujours parfaits (les traductions automatiques sont généralement révisées et corrigées par des professionnel-le-s humains), les systèmes de traduction automatique neuronaux sont très efficaces notamment pour des textes de domaines spécifiques définis (Poibeau, 2020).

Il existe trois types de systèmes de traduction automatique⁸ qui furent développés et utilisés depuis le milieu du XX^{ème} siècle (Poibeau, 2020) :

- Les systèmes à base de règles (1950-1990) qui se basent sur des dictionnaires bilingues (généralement spécialisés) et des règles de transfert (règles grammaticales et linguistiques).
- Les systèmes statistiques (1990-2014) qui se basent sur un traitement statistique de la langue en analysant de grandes quantités de données obtenues grâce à des corpus parallèles.
- Les systèmes neuronaux (2014-) qui se basent sur des réseaux de neurones (modélisation à l'aide de vecteurs du sens des mots).

Aujourd'hui les systèmes de traduction automatique sont généralement neuronaux et leurs performances ne cessent de s'améliorer. Les systèmes de TA sont devenus un des outils du TAL fréquemment utilisés même par les non linguistes, car ils sont faciles d'accès et rapides à utiliser. On peut constater cette popularité en prenant comme exemple le système de traduction automatique Google Translate, un des pionniers de la traduction automatique, qui utilise également le système neuronal depuis 2016. En 2018, on comptait environ 200 millions d'utilisateurs quotidiens de Google Translate (Prates et al., 2018, p. 3).

Néanmoins comme le rappellent Hutchins & Somers (1992, p. 2), l'utilité d'un système de TA est déterminée uniquement par la qualité de ses résultats. Cette qualité peut être mesurée par l'évaluation humaine (qui reste subjective) ou par une évaluation automatique (corrélation avec des traductions humaines). Les problèmes que rencontrent

8 <https://www.sdltrados.com/fr/solutions/machine-translation/>

encore les systèmes de TA sont nombreux et la question du contexte semble être au centre de tous.

3.2. La place du genre dans la traduction automatique

La place du genre et la vision biaisée des systèmes de traduction automatique est un sujet directement relié à l'inclusivité (Vanmassenhove *et al.*, 2018).

Lors d'une traduction, qu'elle soit humaine ou automatique, le contexte étant relié directement à l'auteur du texte source, certains traits spécifiques (comme le genre de l'auteur) vont être perdus (Vanmassenhove *et al.*, 2018, p. 3003). Ce problème est d'autant plus important lors d'une traduction automatique où il a été observé que les erreurs morphologiques d'accord en genre et en nombre sont très nombreuses. De plus, comme le soulignent Vanmassenhove *et al.* (2018, p. 3003) lorsqu'une résolution d'accord a lieu, le masculin est généralement privilégié. Ce résultat s'explique par le fait que généralement les systèmes de TA, contrairement aux humains, ne disposent pas d'un accès au contexte et ainsi vont se baser sur les données sur lesquels leur apprentissage a été réalisé, donc des corpus parallèles. Ces corpus d'entraînement vont alors être la norme sur laquelle vont se reposer les systèmes de traductions automatiques lorsque des questions d'accord (en genre notamment) se posent. Le genre devient ainsi une donnée implicite pour le système de traduction automatique qui, contrairement à l'humain, ne sait pas (encore) gérer cette notion.

De manière générale, c'est le travail même du traducteur ou de la traductrice de savoir gérer les différents rapports au genre selon la langue. Comme l'explique Susie Schmidt (1995), le traducteur ou la traductrice fait des choix en connaissance de ce qui est dit mais aussi de ce qui est implicite. Les traducteurs et les traductrices ne sont pas elles-mêmes toujours d'accord face aux comportements qu'il faut avoir en traitant de préjugés implicites, des a priori ou bien des différences de genre notamment lorsque dans la paire de langues, une des deux langues est non-genrée. Finalement, tandis que le choix de retranscription est personnel (mais doit être justifié), ce qui demeure essentiel pour le traducteur ou la traductrice est de reconnaître les stéréotypes et préjugés implicites qui peuvent découler de la présence (ou non) du genre dans la langue.

3.3. La traduction automatique face aux stéréotypes

La complexité du traitement du genre est directement reliée aux biais genrés qui peuvent renforcer certains stéréotypes et certaines discriminations. De manière générale, cela pose problème car comme nous l'avons déjà montré précédemment, le langage peut affecter nos perceptions et représentations du monde.

Prates *et al.* (2018) ont réalisé une étude détaillée des problèmes de traduction du système de TA Google Translate. L'étude se concentre sur le traitement des noms de métiers et de quelques adjectifs en prenant comme langue source l'anglais, et on observe des traductions privilégiant les pronoms masculins aux pronoms féminins ou neutres

même si les langues cibles choisies (comme l'anglais, le turc, le basque, l'estonien ou le hongrois) sont toutes des langues possédant des structures alternatives neutres. Selon le domaine, le pourcentage de traductions préférant le genre masculin pour les pronoms va varier avec des valeurs très élevées pour les domaines scientifiques et informatiques ce qui correspond aux stéréotypes connus. L'étude montre également la place des traductions féminines des pronoms désignant des noms de métiers. On retrouve surtout des traductions féminines pour les noms de métiers des domaines de la santé et de l'éducation. De manière plus précise, l'étude nous montre également que certaines langues sont plus ou moins touchées par ces traductions à tendance masculine, avec par exemple les traductions vers le malais ou en estonien qui ont un pourcentage, de traductions utilisant les pronoms masculins supérieur à 70 %. D'autres langues au contraire, comme le basque ou le yoruba, ont des pourcentages de traductions utilisant le neutre, supérieur à 30 %. L'étude s'est également intéressée au traitement des adjectifs. Lors de la traduction automatique de certains adjectifs par Google Translate. Prates *et al.* (2018) ont relevé que lorsqu'une phrase contient des adjectifs tels que « timide » (*shy*) ou « gentil » (*kind*), le genre utilisé de manière générale sera le féminin. Pour des phrases contenant des adjectifs comme « arrogant » (*arrogant*) ou « coupable » (*guilty*), le genre utilisé de manière générale sera le masculin.

Ungerleider (2013) reprend les travaux réalisés par Londa Scheibinger qui montrait déjà en 2013 les problèmes de biais dans les systèmes de TA tels que Google Translate ou Systran. Scheibinger montrait notamment que lorsqu'elle avait cherché à traduire l'un de ses articles où l'on retrouvait des formes non-genrées espagnoles tels que « dice » (« il/elle dit ») ou « suyo » (« son/sa ») les systèmes de TA Google Translate et Systran utilisait le genre masculin en anglais (« he says », « his »). Ces problèmes de traduction ne sont pas liés à une langue ou à un couple de langues mais bien au système de TA comme le montre Ungerleider avec des exemples de travaux sur l'espagnol, l'allemand ou l'hébreu.

L'une des raisons qui explique le biais vers le genre masculin du système de TA réside dans le choix des données servant à entraîner les systèmes de TA (Ungerleider, 2013, p. 1). De plus, les systèmes de TA (notamment les systèmes statistiques) se basent généralement sur le résultat le plus probable (en mesurant la fréquence) ainsi les résultats masculins (ou neutres mais ayant une forme identique au masculin) vont être plus fréquents (Ungerleider, 2013, p. 3).

Ces différentes études nous permettent donc de mettre en avant le problème du genre à l'intérieur même des outils de traitement automatique du français, avec des stéréotypes qui sont assimilés et renforcés. Concernant Google Translate, en décembre 2018, une nouvelle fonctionnalité a été ajoutée afin de proposer une traduction à la fois masculine et féminine (Prates *et al.*, 2018, p. 27). Ce changement montre qu'il est possible de modifier et d'améliorer les outils du TAL afin de réduire la tendance à la masculinisation. Ces progrès ne sont cependant pas équivalents pour toutes les langues et ne permettent pas d'inclure la non-binarité (Prates *et al.*, 2018, p. 27). Une autre solution est de pencher

vers une écriture plus inclusive ce qui va permettre d'introduire plus de contextes où les genres seront différents, et non stéréotypés, afin de réduire les problèmes actuels. Cependant cette écriture inclusive de par ses différentes formes (souvent nouvelles) pourrait poser problème aux systèmes de TA et ainsi ne pas permettre d'obtenir des traductions pertinentes. Nous avons ainsi décidé de mener une étude sur le traitement de l'écriture inclusive par les systèmes de TA.

4. ÉTUDE DU COMPORTEMENT DES SYSTÈMES DE TRADUCTION AUTOMATIQUE NEURONAUX FACE À L'ÉCRITURE INCLUSIVE

Notre étude portera sur le comportement des systèmes de traduction automatique neuronaux face à l'écriture inclusive. Nous allons traduire les textes d'un corpus français en anglais, puis par la suite réaliser une traduction inversée (traduction en français) et observer tout le long les diverses transformations et changements réalisés. Nos questions de recherche durant cette étude seront :

- A quoi ressemblent les formes d'écriture inclusive tout au long de notre étude (selon la langue et le système de traduction automatique) ?
- Comment les systèmes de traduction automatique sont-ils affectés par l'écriture inclusive ?
- Y a-t-il un rapport entre les stratégies discursives (formes d'écriture inclusive) et les stratégies de traduction (comportement des TA face à l'écriture inclusive) ?

4.1. Méthodologie

Dans un premier temps, nous avons récolté des données textuelles en formant notre propre corpus à partir d'articles de journaux et magazines en ligne afin d'avoir des thèmes variés mais surtout afin d'étudier des textes où l'écriture inclusive n'est pas utilisée par des spécialistes. Afin de trouver ces articles, nous avons réalisé une première recherche pour voir si certains journaux ou magazines avaient indiqué leur utilisation de l'écriture inclusive. Certains, comme le Monde, ont en effet exprimé leur soutien envers l'écriture inclusive ou du moins la volonté d'une utilisation plus répandue par le futur. Cela nous a permis de trouver deux premiers articles venant de sites encourageant la féminisation : un premier article (Texte 1) du magazine en ligne Madmoiselle portant sur le choix d'un ou une psychologue (Justine, 2020) et un deuxième article (Texte 2) du magazine en ligne Slate portant le travail en équipe (Boukobza, 2020).

Tandis que nous cherchions des articles utilisant l'écriture inclusive, traitant d'un sujet autre que la linguistique, nous avons trouvé le groupe Facebook Taglinclusive⁹ qui est un groupe public ayant comme volonté de « référencer les pratiques de langage et d'écriture inclusive pour montrer que cette manière d'écrire et de s'exprimer est désormais répandue. ». Grâce à ce groupe, nous avons pu trouver deux articles : le premier (Texte 3) portant sur les violences policières dénoncées par les journalistes sur le site du journal FranceInfo (Journalistes ayant signé la tribune, 2019) et le deuxième (Texte 4) portant sur l'élection de la Première Ministre de Finlande sur le site du journal RFI (RFI, 2019). Ce groupe Facebook nous a aussi permis de repérer d'autres sites qui utilisent ou ont utilisé cette écriture comme le site Libération. Grâce à une recherche thématique (portant sur les étudiant·e·s, un des sujets où l'on retrouve fréquemment l'utilisation de l'écriture inclusive

9 Groupe Facebook Taglinclusive : <https://www.facebook.com/groups/208287676407405>

notamment avec les doubles formes abrégés), nous avons trouvé l'article (Texte 5) du journal Libération portant sur le sexisme, l'homophobie et le racisme dans les grandes écoles de commerce (Un collectif d'ancien·ne·s et actuel·le·s étudiant·e·s des grandes écoles de commerce françaises, 2020).

Afin d'avoir un corpus assez complet, nous avons par la suite cherché sur la base de données d'informations Europresse¹⁰ des exemples concrets des différentes stratégies discursives (« iel », « Français et Françaises », « elles et ils ») afin d'avoir une représentation plus équilibrée des différentes stratégies. Ces recherches nous ont permis d'obtenir nos trois derniers articles : un article (Texte 6) portant sur le secret médical sur le site du journal HuffingtonPost (Winckler, 2019), un article (Texte 7) portant sur la prise de poids pendant le 1^{er} confinement sur le site du journal 20minutes (Garrat-Valcarcel, 2020) et enfin un article (Texte 8) traitant du sujet du RSA pour les moins de 25 ans, sur le site du journal de l'Humanité (Tribune collective, 2020). Les articles sont tous relativement récents (5 articles rédigés en 2020 et 3 articles rédigés en 2019) afin d'avoir des utilisations d'écriture inclusive actuelles. Nous nous retrouvons ainsi avec un corpus de 8 articles. Afin d'obtenir le nombre approximatif de tokens nous avons utilisé l'outil TreeTagger¹¹. Notre total est de 10 645 tokens et le Tableau 3 détaille le nombre de tokens par article.

Article : source (date)	Nombre de tokens
Texte 1 : Madmoizelle (2020)	1 888
Texte 2 : Slate (2020)	500
Texte 3 : Franceinfo (2019)	1 304
Texte 4 : RFI (2019)	281
Texte 5 : Libération (2020)	2 726
Texte 6 : Huffingtonpost (2019)	2 864
Texte 7 : 20minutes (2020)	382
Texte 8 : l'Humanité (2020)	700
Total	10 645

Tableau 3 : Nombre de tokens par article

Une fois notre corpus formé, nous avons annoté ces articles à l'aide de l'annotateur Glozz¹² afin d'identifier les différentes stratégies discursives utilisées. Le schéma d'annotation que nous avons utilisé avec Glozz (Figure 1) est basé sur une classification en cinq différentes stratégies discursives bien que d'autres classifications ont été proposées comme les trois conventions introduites par Sebahg et Barric (2019).

Les cinq stratégies discursives sont :

¹⁰ Base de données d'informations : <http://www.europresse.com/fr/>

¹¹ Étiquetage morpho-syntaxique, lemmatisation et: outil de tokenisation
<https://cental.uclouvain.be/treetagger/>

¹² Plateforme d'annotation : <http://www.glozz.org>

- **Doubles formes abrégées** : forme utilisant une structure composite avec la forme masculine, un symbole (point médian, parenthèse, barre oblique, tiret), puis le suffixe féminin.
- **Doubles formes intégrales**: utilisation des deux formes (forme féminine et forme masculine), l'une à côté de l'autre afin de renvoyer à une seule et même idée.
- **Règle de proximité** : accord avec le genre du référent le plus proche et non pas une résolution à l'accord au genre masculin.
- **Créations langagières (néologismes et féminisation)**: nouveaux mots et nouveaux termes créés récemment et encore peu usités, dans le but d'avoir des équivalents féminins ou non-genrés.
- **Épicènes et formulations englobantes**: utilisation d'épicènes (forme sans genre intrinsèque) et formulations génériques (pluriel ou forme non-individualisante).

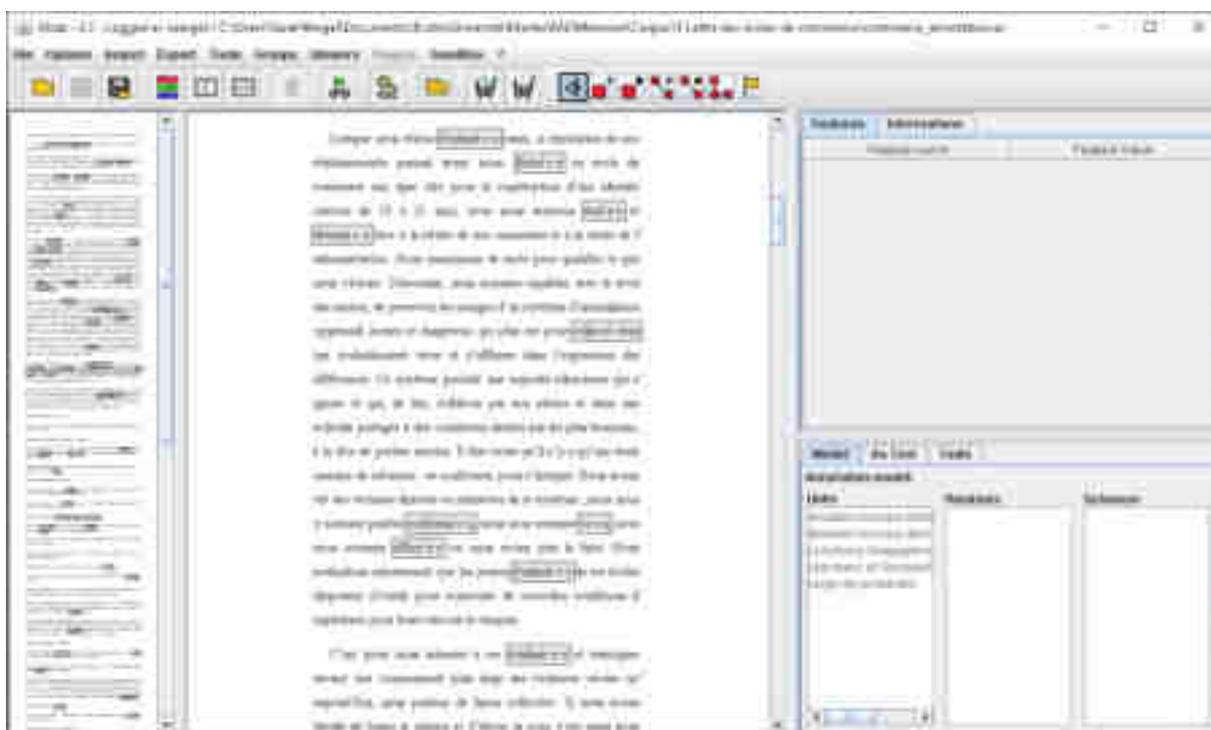


Figure 1 : Capture d'écran de l'interface d'annotation des stratégies discursives en français de la plateforme Glozz

Sur les 8 articles de notre corpus, nous avons un total de 249 exemples d'écriture inclusive annotés. Le graphique (Figure 2) fournit le nombre d'exemples présents dans le corpus pour chaque stratégie discursive d'écriture inclusive.

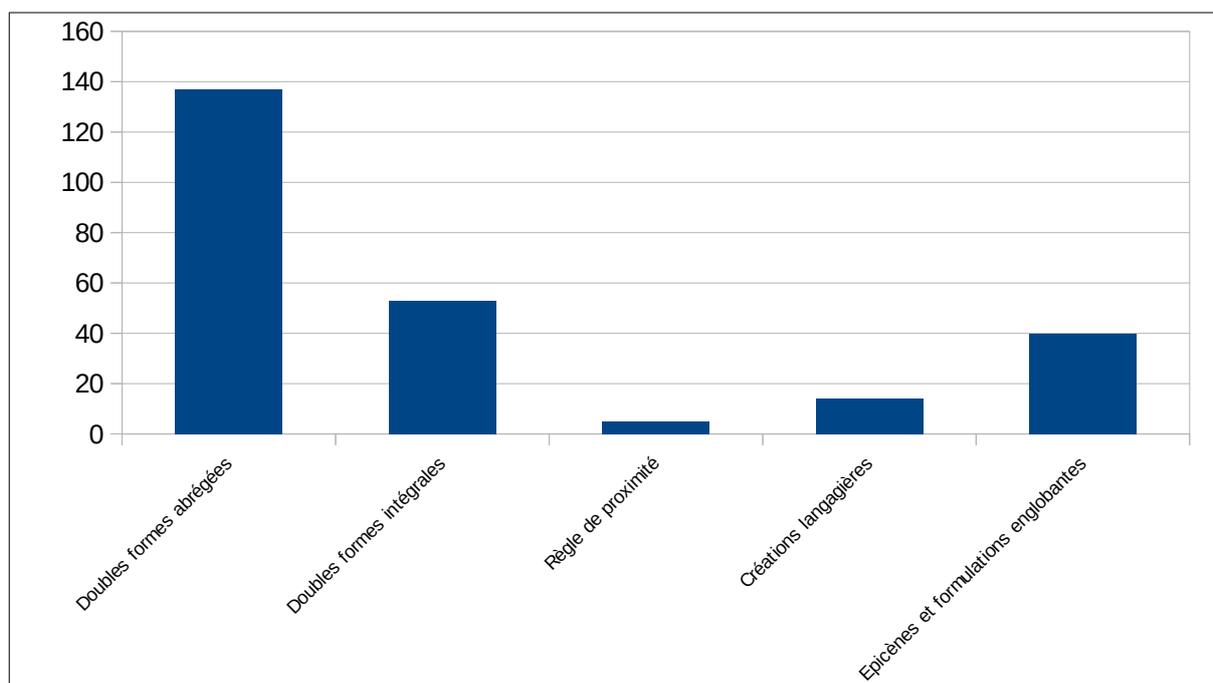


Figure 2 : Nombre d'occurrences des différentes stratégies discursives d'écriture inclusive

Ayant tenté d'obtenir des exemples des cinq formes d'écriture inclusive, nous avons cherché des textes variés dans leur utilisation de l'écriture inclusive, néanmoins, on remarque tout de même une faible présence d'exemples d'utilisation de la règle de proximité ou de créations langagières. Ces deux stratégies semblent être les moins utilisées, ou les moins répandues dans des textes journalistiques.

Une fois les textes annotés, nous avons traduit les textes français en anglais à l'aide des systèmes de traduction automatique neuronaux. Ces traductions vont nous permettre tout le long de notre étude d'identifier les problèmes que peuvent poser les exemples d'écriture inclusive, mais aussi nous permettre d'analyser les stratégies empruntées par les systèmes de TA pour traduire adéquatement les différentes stratégies discursives. Pour cela, nous avons choisi trois systèmes de traduction automatique neuronaux, qui sont les plus performants et populaires sur le marché actuellement (en plus d'être facile d'utilisation et d'accès) : DeepL¹³, Google Translate¹⁴ et Systran¹⁵.

DeepL est un système de traduction automatique mis en ligne en août 2017¹⁶. Il se démarque par la qualité des traductions proposées grâce à l'utilisation de réseaux de neurones¹⁷. DeepL est disponible pour 26 langues (mars 2021¹⁸) et il est également possible depuis 2020 de sélectionner la traduction la plus appropriée selon l'utilisateur

13 Système de traduction automatique DeepL : <https://www.deepl.com/fr/translator>

14 Système de traduction automatique GoogleTranslate : <https://translate.google.fr/?hl=fr>

15 Système de traduction automatique Systran : <https://translate.systran.net/translationTools/text>

16 <https://www.deepl.com/press.html>

17 <https://www.deepl.com/fr/blog/20200206/>

18 <https://www.deepl.com/blog/20210316.html>

(fonctionnalité « glossaire »). Au service gratuit, s'ajoute le service payant DeepL Pro qui permet d'utiliser l'API de DeepL.

Google Translate est un système de traduction automatique mis en ligne en avril 2006. Utilisant d'abord la traduction statistique, le système passe aux réseaux de neurones en novembre 2016¹⁹. Google Translate est disponible pour 108 langues (Février 2020²⁰) et sa version mobile permet d'obtenir une traduction instantanée en pointant son appareil photo sur du texte, ou en parlant directement dans le micro²¹. Au service gratuit, s'ajoutent également les API payantes Translation qui permettent d'utiliser les modèles de traduction pré-entraînés de Google²².

L'entreprise Systran (SYStem TRANslation), fondée en 1968, est l'une des plus anciennes entreprises de traduction automatique. Le moteur de traduction lancé en 2009 s'appuie sur un hybride entre traduction statistique et à base de règles²³, puis en 2016 le moteur de traduction Pure Neural MT marque le passage aux réseaux de neurones²⁴. Systran Translate est disponible pour 55 langues (mai 2021²⁵). Au service gratuit (mais limité), s'ajoute également Systran Translate PRO qui permet d'accéder à l'API de traduction, de personnaliser les traductions et de traduire des documents de tout type de format.

En utilisant le format textuel de chaque article, nous avons pu réaliser les traductions sur les différents sites des systèmes de traduction automatique, puis nous avons récupéré ces traductions sous format textuel. Les traductions des huit articles ont été réalisées le 11 novembre 2020. Nous fournissons la date, car des changements sont fréquemment réalisés sur ces outils.

Afin d'analyser ces traductions, nous avons à nouveau utilisé un annotateur pour observer directement les stratégies utilisées par les systèmes de TA. A l'aide de l'outil Glozz, nous avons annoté les traductions des exemples d'écriture inclusive selon la stratégie de traduction utilisée (Figure 3). Nous avons décidé pour cela de choisir les trois stratégies de traduction les plus utilisées par les traducteurs, c'est-à-dire :

- **Omission** : suppression de la forme dans la traduction (dans notre cas, cela correspond à l'absence de forme d'écriture inclusive)
- **Remplacement** : changement de la forme (dans notre cas, cela correspond à un changement par une neutralisation ou masculinisation de la forme d'écriture inclusive)
- **Maintien** : forme identique (ou très similaire) dans la traduction (dans notre cas, cela correspond à la présence de la même forme d'écriture inclusive)

19 <https://www.independent.co.uk/life-style/gadgets-and-tech/news/how-does-google-translate-work-b1821775.html>

20 <https://www.blog.google/products/trans/>

21 <https://translate.google.com/about/?hl=fr>

22 <https://cloud.google.com/translate/?hl=fr#section-1>

23 <https://bfmbusiness.bfmtv.com/01-business-forum/de-la-guerre-froide-a-la-coree-l-incroyable-epopee-de-systran>

24 <https://www.actuia.com/actualite/traduction-automatique-focus-sur-la-solution-open-source-proposee-par-systran-acteur-historique-de-la-t-a-qui-ne-cesse-de-se-renouveler/>

25 <https://www.systransoft.com/fr/>

Bien que ces stratégies soient généralement plutôt utilisées pour traduire les créations langagières ou les noms propres (Leppihalme, 1997), nous avons remarqué, après avoir examiné plusieurs typologies de stratégies de traduction (Owji, 2013) que de manière générale ces trois stratégies sont les plus usitées (sous diverses variantes) peu importe la typologie. À ces trois stratégies, nous avons néanmoins rajouté une stratégie de traduction, **l'ajout**, car nous avons remarqué après les premières traductions, que les système de TA avaient parfois tendance à rajouter des formes d'écriture inclusive (non présentes dans le texte original français) pour deux raisons. Ces ajouts étaient dus à une volonté de cohérence, notamment lors de la traduction du possessif, ou ils étaient dus à une nécessité de réintroduire l'inclusivité là où une précédente forme d'écriture inclusive avait été transformée (en raison d'une neutralisation) ou bien omise.

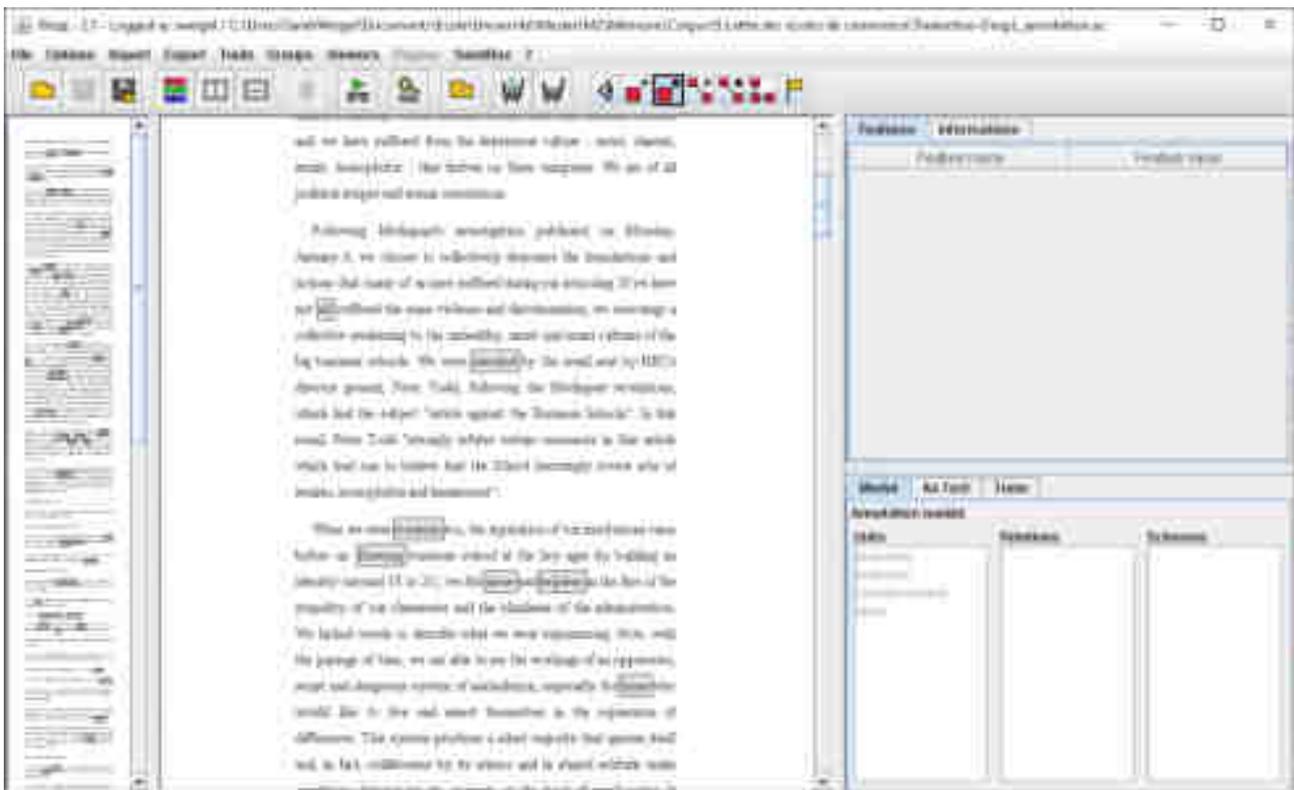


Figure 3 : Capture d'écran de l'interface d'annotation des stratégies de traduction de la plateforme Glozz

La typologie que nous avons construite a pour unique but d'analyser le comportement des systèmes de TA face aux exemples d'écriture inclusive. Ainsi, on ne s'interroge pas sur la qualité même de la traduction (correcte ou incorrecte) bien que certaines erreurs de traduction ont pu être relevées dans notre analyse. De plus, on ne tient pas compte des particularités langagières de la langue source. Ainsi, ici, bien que l'anglais soit une langue non genrée, on ne cherche pas à observer les possibilités de traduction d'un point de vue linguistique, mais plutôt d'un point de vue technique, c'est-à-dire comment le système de TA va traiter les exemples d'écriture inclusive. Le Tableau 4 reprend les quatre stratégies de traduction avec des exemples obtenus après l'analyse de nos résultats.

Stratégie de traduction	Avant traduction	Après traduction
Omission	« Nous aussi, nous nous sommes tu·e·s » (Texte 5)	« We too have · and · » (Traduction Systran)
Remplacement	« Toutefois, la professeure rappelle que la mixité bonifie le travail d'équipe » (Texte 2)	« However, the professor reminds us that mixing improving teamwork » (Traduction Systran)
Maintien	« La personne qui demande des soins n'est pas un cobaye sur laquelle on peut agir arbitrairement. » (Texte 6)	« The person seeking care is not a guinea pig on which we can act arbitrarily. » (Traduction Google Translate)
Ajout	« Ensuite, chaque praticien ou praticienne aura sa méthode, son courant de passé... » (Texte 1)	« Then, each practitioner will have his or her own method, his or her own current of thought ... » (Traduction DeepL)

Tableau 4 : Exemples d'utilisation des stratégies de traduction

Afin d'avoir un meilleur accès aux comportements des systèmes de TA face aux différentes stratégies discursives, nous avons réalisé un jeu de données annoté complet constitué des phrases parallèles avec leurs annotations. Ce jeu de données met en parallèle le texte original (français) avec les traductions des trois systèmes (anglais).

Nous avons développé un script python²⁶ permettant d'annoter automatiquement les textes de notre corpus avec les étiquettes correspondantes. Nous avons utilisé les fichiers annotés (fichier .aa) avec Glozz afin de récupérer les indices des unités annotés, ainsi que les fichiers textuels (fichier .ac). Pour les textes français, nous avons annoté les stratégies discursives (doubles formes abrégées, doubles formes intégrales, règle de proximité, créations langagières, épïcènes et formulations englobantes) avec les étiquettes suivantes :

- DFA (doubles formes abrégées)
- DFI (doubles formes intégrales)
- RP (règle de proximité)
- CL (créations langagières)
- EFE (épïcènes et formulations englobantes)

²⁶ https://git.unistra.fr/weigels/ecritures_inclusives_ta

Pour les textes anglais (traductions des trois systèmes de traduction automatique), nous avons annoté les stratégies de traductions (omission, maintien et remplacement) avec les étiquettes suivantes :

- O (omission)
- M (maintien)
- R (remplacement)

Nous n'avons pas annoté les ajouts car ces formes ne peuvent pas être rattachées à des formes de stratégies discursives dans les textes d'origine.

Nous obtenons ainsi 4 fichiers (1 français, 3 anglais) sous format txt avec les formes annotées. Nous avons ensuite réalisé une vérification et correction manuelle de ces fichiers.

Afin de pouvoir comparer les différentes traductions obtenues pour chaque texte du corpus selon le système de TA, nous avons aligné nos fichiers en utilisant l'outil d'alignement automatique Web Align Toolkit (WAT)²⁷. Nous avons obtenu un fichier csv²⁸ avec le texte original (français) et les trois traductions (anglais). Enfin, nous avons réalisé une nouvelle correction manuelle car certaines phrases n'avaient pas été séparées correctement.

4.2. Analyse des formes d'écriture inclusive après traduction automatique

4.2.1. Utilisation des stratégies de traduction par les systèmes de traduction automatique

Après que les traductions ont été obtenues et annotées, nous pouvons analyser le comportement des systèmes de TA face à l'écriture inclusive et la manière dont les stratégies de traduction (vues précédemment) ont été utilisées notamment selon la stratégie discursive.

Pour notre premier système de traduction, DeepL, on peut commencer par noter que les problèmes posés par l'écriture inclusive et ses stratégies discursives sont plutôt minimes notamment par rapport à ce que nous avons supposé.

Dans le Tableau 5, nous pouvons voir quelles ont été les stratégies de traduction les plus utilisées par le système de TA DeepL selon la stratégie discursive.

²⁷ <http://phraseotext.univ-grenoble-alpes.fr/webAlignToolkit/>

²⁸ https://git.unistra.fr/weigels/ecritures_inclusives_ta/-/tree/master/Alignement_clean

	Doubles formes intégrales	Double formes abrégées	Règle de proximité	Créations langagières	Épicènes et formulations englobantes
Remplacement	37	137	1	10	0
Omission	0	0	4	2	2
Maintien	16	0	0	2	38

Tableau 5 : Stratégie de traduction selon la stratégie discursive pour le système de TA DeepL

Les doubles formes abrégées ont été bien reconnues et n'ont pas posé de problèmes par leur structure composite (notamment par rapport à la présence de symboles tels que les points médians, tirets ou parenthèses). Les doubles formes abrégées ont été généralement remplacées par des formes neutres (telles que des épïcènes) ce qui s'explique par le fait que l'anglais n'est pas une langue grammaticalement genrée (contrairement au français) ainsi des alternatives neutres existent. On retrouve par exemple, l'utilisation de « those » et « all » :

Exemple 1:

Texte 5 : « Ce “ nous ” n'est pas homogène, mais il veut rassembler **tout·e·s celles et ceux** qui ont subi ces agressions (...) »

Traduction DeepL : « This “ we ” is not homogenous, but it wants to bring together **all those** who have suffered these attacks (...) »

On pourrait dire que ce remplacement est efficace du point de l'écriture inclusive, car il assure une plus grande neutralité (notamment au-delà de la binarité du genre) cependant cela revient tout de même à retirer l'effort de représentation féminine réalisé en français. Les doubles formes intégrales sont généralement maintenues lorsqu'elles sont brèves et qu'il est possible de traduire la dualité des genres telles que « il/elle » ou « il et elle » traduit par « he/she » :

Exemple 2 :

Texte 1 : « Vous pouvez également l'interroger sur ce qu'**il ou elle** propose “ en cas d'urgence ”, ou lors de ses absences – par exemple, peut-**il/elle** vous référer à un confrère si vous avez une urgence lors de ses congés ? »

Traduction DeepL : « You can also ask him/her what **he/she** proposes “in case of emergency” , or during **his/her** absences – for example, can **he/she** refer you to a colleague if you have an emergency during **his/her** vacations ? »

Lorsque deux formes genrées n'existent pas en anglais, les doubles formes intégrales sont remplacées par des formes neutres (tels que des épïcènes) comme « praticien ou praticienne » traduit par « practitioner » :

Exemple 3 :

Texte 1 : « Je choisis le "type" de **praticien ou praticienne** »

Traduction DeepL : « I choose the "type" of **practitioner** »

Lorsque les doubles formes intégrales sont remplacées, on retrouve des ajouts de formes d'écriture inclusive telles que « his or her » afin de maintenir une certaine inclusivité de l'écriture (réintroduire les deux genres) :

Exemple 4 :

Texte 1 : « Ensuite, chaque **praticien ou praticienne** aura sa méthode, son courant de pensée ... »

Traduction DeepL : « Then, each **practitioner** will have **his or her** own method, **his or her** own current of thought... »

Il arrive cependant que les doubles formes intégrales soient maintenues et dans ce cas-là, elles posent souvent problème. Le maintien ne va pas fonctionner en anglais et on va avoir une répétition comme par exemple « Français et Française » qui est traduit par « French and French people » :

Exemple 5 :

Texte 7 : « Finalement, le confinement, qui approche des deux mois en France, n'aura pas tant fait grossir que ça **Français et Française**. »

Traduction DeepL : « In the end, the confinement, which is approaching two month in France, will not have made **French and French people** fatter than that. »

Concernant l'utilisation de la règle de proximité, ces formes sont toujours omises car l'accord en genre n'existe pas en anglais comme c'est une langue grammaticalement non-genrée :

Exemple 6 :

Texte 7 : « Choisir **un ou une psy qui vous convienne** n'est pas forcément évident, et surtout, n'arrive pas toujours du premier coup. »

Traduction DeepL : « Choosing **the right shrink for you** is not always easy, and above all, it doesn't always happen the first time. »

Les créations langagières sont remplacées par des formes neutres existant déjà en anglais, ou bien par des tournures différentes permettant l'inclusivité. On retrouve par exemple le pronom personnel neutre « they » qui remplace le pronom personnel français « iels » :

Exemple 7 :

Texte 6 : « *De plus, comme être médecin n'évite nullement collusion et conflits d'intérêts, il n'est pas sûr que tous les médecins dénonceraient des hommes qui brutalisent leur compagne... mais avec qui **iels** ont l'habitude de déjeuner ou de jouer au golf ou à la pétanque. »*

Traduction DeepL : « *Moreover, since being a doctor does not avoid collusion and conflicts of interest, it is not certain that all doctors would denounce men who brutalize their partners, but with whom **they** usually have lunch or play golf or petanque »*

Enfin pour les épïcènes et les formulations englobantes, on a généralement des maintiens, car l'anglais utilise des équivalents directs comme par exemple « membres » traduit par « members » :

Exemple 8 :

Texte 2 : « *Bien que les **membres** ne puissent pas se voir ni appréhender les expressions faciales des autres, les **personnes** considérant sincèrement les propos de leurs **collègues** et leur façon de s'exprimer sont un véritable atout. »*

Traduction DeepL : « *Although **members** cannot see each other or understand each other's facial expressions, **people** who sincerely consider what their **colleagues** say and how they express themselves are a real asset. »*

Certaines omissions d'épïcènes pourraient s'expliquer par la volonté de fluidifier la style, car il arrive que ces formes deviennent répétitives en anglais. De manière générale, l'utilisation de formulations englobantes est très fréquente afin d'éviter toute spécification genrée ou individualisante en anglais.

Nous retrouvons ci-dessous (Figure 4) une représentation de la répartition de l'utilisation des stratégies de traduction (omission, remplacement, maintien, ajout) au sein de notre corpus traduit par DeepL.

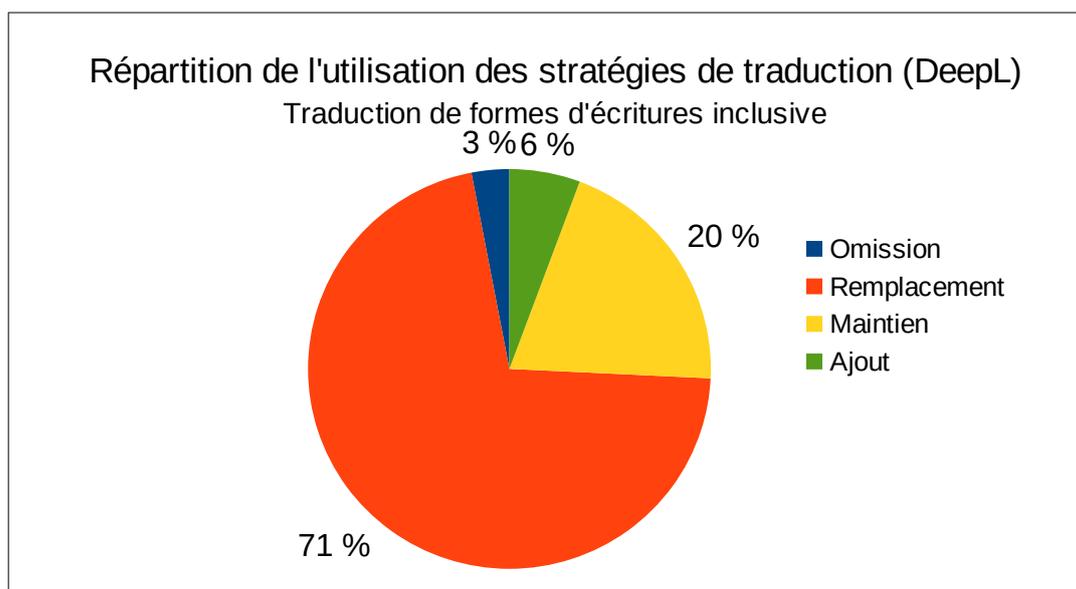


Figure 4 : Répartition de l'utilisation des stratégies de traduction pour le système de TA DeepL

Pour le deuxième système de traduction, Google Translate, de nouveaux problèmes ont été observés notamment en lien avec les doubles formes abrégées.

Dans le Tableau 6, nous pouvons voir quelles ont été les stratégies de traduction les plus utilisées par le système de TA Google Translate selon la stratégie discursive.

	Doubles formes intégrales	Double formes abrégées	Règle de proximité	Créations langagières	Épicènes et formulations englobantes
Remplacement	35	137	1	9	0
Omission	0	0	4	2	0
Maintien	18	0	0	3	40

Tableau 6 : Stratégie de traduction selon la stratégie discursive pour le système de TA Google Translate

Les doubles formes abrégées vont être généralement remplacées par des formes neutres (telles que des épïcènes) car comme nous l'avons déjà précisé, l'anglais est une langue grammaticalement non-genrée. Cependant, certaines doubles formes abrégées vont poser problème au système de traduction automatique comme par exemple « nombreux·ses » en raison du suffixe « ses » qui va être traité comme le pronom possessif pluriel par le système de traduction automatique et ainsi traduit par « his » :

Exemple 9 :

Texte 3 : « *Nous sommes de **nombreux·ses** journalistes à nous en plaindre.* »

Traduction Google Translate : « *We are **many · his** journalists to complain about it.* »

Ces mauvais traitements des morphèmes des doubles formes abrégées sont récurrents avec parfois un maintien du morphème dans la traduction, qui est juxtaposé à la forme traduite. De manière plus spécifique, certaines formes précises de doubles formes abrégées ont rendu la traduction erronée comme « tu·e·s » (du verbe « taire »). Cette forme a été reconnue comme son homonyme « tue » (du verbe « tuer ») et on retrouve ainsi le terme « killed » dans la traduction proposée par Google Translate :

Exemple 10 :

Texte 5 : « *Nous nous sommes **tu·e·s** par manque de moyens.* »

Traduction Google Translate : « *We **killed** each other for lack of resources.* »

Ici, en plus de l'homonymie, le point médian pose problème. Enfin, concernant toujours les doubles formes abrégées, l'utilisation de la structure composite des doubles formes abrégées sur un anglicisme, ici le terme « out », va poser problème au système de traduction automatique qui n'a pas su traiter la structure « outé·e·s » et l'a remplacé par « outcry » :

Exemple 11 :

Texte 5 : « *Nous aussi, nous avons eu peur de sortir du placard, ou nous avons dû y rentrer à nouveau, quand nous n'avons pas été «**outé·e·s**» sans notre consentement.* »

Traduction Google Translate : « *We too were afraid to come out of the closet, or had to come back in again, when we were not "**outcry**" without our consent.* »

Généralement les doubles formes intégrales sont remplacées par des formes neutres. Il arrive aussi que certaines doubles formes intégrales soient maintenues comme « son ou sa » qui est traduit par « his or her » même si dans l'exemple suivant le sens change lors de la traduction. Dans la phrase d'origine, les pronoms possessifs renvoient à la personne lisant le texte (choisir un ou une psy pour soi-même), tandis que dans la traduction, les pronoms possessifs renvoient à une autre personne (choisir un ou une psy pour quelqu'un d'autre) :

Exemple 12 :

Texte 1 : « *Comment choisir **son ou sa** psy ?* »

Traduction Google Translate : « *How to choose **his or her** shrink?* »

Dans d'autres cas, le maintien n'ajoute pas de l'inclusivité, compliquant au contraire le texte :

Exemple 13 :

Texte 1 : « Pour dégoter **un ou une** chouette psy... »

Traduction Google Translate : « To find **one or a nice** shrink ... »

Ce problème de maintien erroné est fréquent avec parfois des répétitions, car la traduction sera neutre et ainsi identique, par exemple « futur ou future » traduit par « future or future » :

Exemple 14 :

Texte 1 : « Le premier contact avec votre **futur ou future** thérapeute se fera probablement par téléphone. »

Traduction Google Translate : « The first contact with your **future or future** therapist will probably be by phone. »

Ce problème n'apparaît cependant pas pour toutes les doubles formes intégrales ce qui peut laisser penser que c'est lié à un apprentissage des formes les plus courantes car on a aucun problème par exemple pour « étudiants et étudiantes » qui est traduit par « students » :

Exemple 15 :

Texte 8 : « (...) le versement d'une aide exceptionnelle de 200 euros pour les précaires de moins de 25 ans, ainsi que pour les **étudiantes et étudiants** ayant perdu leur emploi. »

Traduction Google Translate : « (...) the payment of exceptional aid of 200 euros for precarious under 25, as well as for **students** who lost their jobs. »

Les exemples d'utilisation de la règle de proximité sont tous omis en anglais, car il n'y a pas d'accord en genre. Les créations langagières sont soit omises soit remplacées par des formes neutres (tels que les épécènes), car l'anglais ne réalise pas de distinction genrée sur les noms de métiers ou de statut comme « Première ministre » qui est traduit par « Prime Minister » :

Exemple 16 :

Texte 4 : « Il s'agira donc d'un premier test pour la jeune **Première ministre** qui veut prouver que son jeune âge ne sera pas un obstacle. »

Traduction Google Translate : « It will therefore be a first test for the young **Prime Minister** who wants to prove that her young age will not be an obstacle. »

Cependant pour la création langagière « iel » (pronom personnel neutre), un remplacement a lieu avec le pronom masculin « he » :

Exemple 17 :

Texte 6 : « *Le médecin est dépositaire des confidences, **iel** n'a pas le droit de les utiliser à sa guise. »*

Traduction Google Translate : « *The doctor is the custodian of the confidences, **he** does not have the right to use them as he pleases. »*

Cette masculinisation pourrait s'expliquer par le fait que dans la même phrase, le terme « médecin » est utilisé et bien que ce soit un nom épïcène, il garde une forme plutôt masculine. Néanmoins dans une autre phrase plus loin, on relève la même masculinisation avec le pronom « he » bien que les formes genrées présentes dans la phrase d'origine en français étaient des doubles formes abrégées ou intégrales telles que « le/la » :

Exemple 18 :

Texte 6 : « *Ensuite seulement contre **le/la professionnel.le concerné.e**, qui - outre le risque de représailles qu'**iel** encourrait à son tour - serait **mis.e** dans l'impossibilité de soigner de nouveau cette femme si elle en a besoin, (...) »*

Traduction Google Translate : « *Then only against **the professional concerned**, who - in addition to the risk of reprisals that **he** would incur in turn - would **be made** unable to treat this woman again if she needed it »*

Encore un peu plus loin dans le texte, la forme « iel » est correctement traitée, car elle a été traduite par l'équivalent anglais neutre « they » :

Exemple 19 :

Texte 6 : « *Et c'est parce qu'**iel** ne sait pas tout qu'**iel** doit la respecter et ne pas intervenir de manière inconsidérée, pour, d'abord, ne pas nuire. »*

Traduction Google Translate : « *And it is because **they** do not know everything that **they** must respect it and not intervene in a reckless manner, in order, first of all, not to harm. »*

Ces exemples montrent une certaine incohérence de traitement de cette création langagière. Il faut également noter que la forme pluriel « iels » est correctement traduite en anglais par « they » :

Exemple 20 :

Texte 6 : « *(...) mais avec qui **iels** ont l'habitude de déjeuner ou de jouer au golf ou à la pétanque. »*

Traduction Google Translate : « *(...) but with whom **they** are used to having lunch or playing. golf or pétanque. »*

Enfin, concernant les épïcènes et les formulations englobantes, aucun problème n'apparaît, car elles sont maintenues grâce aux équivalents neutres anglais, par exemples « les moins de 25 ans » est traduit par « those under 25 » :

Exemple 21 :

Texte 8 : « *Pourtant, **les moins de 25 ans** n'ont pas accès au revenu de solidarité active (RSA).* »

Traduction Google Translate : « *However, **those under 25** do not have access to active solidarity income (RSA).* »

L'utilisation d'ajouts est toujours présente en partie grâce à l'utilisation de doubles formes intégrales en anglais tels que « his/her » :

Exemple 22 :

Texte 1 : « *De votre côté, si vous ne l'avez pas fait lors de votre premier appel, vous avez toujours la liberté de **le/la** questionner sur sa formation, sa spécialisation, sa méthode.* »

Traduction Google Translate : « *For your part, if you did not do it during your first call, you always have the freedom to question **him / her** about **his / her** training, **his / her** specialization, **his / her** method.* »

Nous retrouvons ci-dessous (Figure 5) une représentation de la répartition de l'utilisation des stratégies de traduction (omission, remplacement, maintien, ajout) au sein de notre corpus traduit par Google Translate.

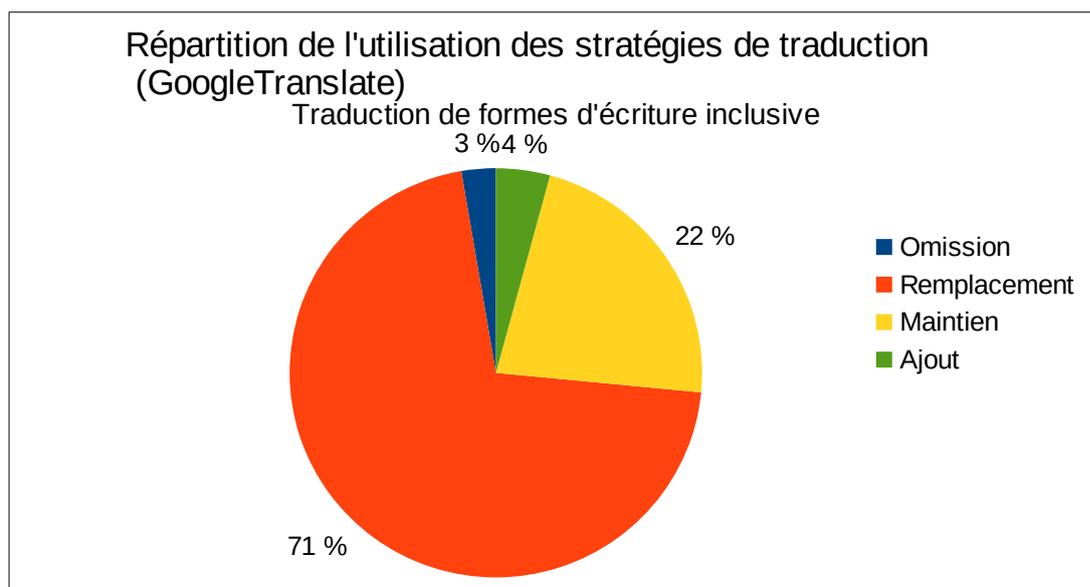


Figure 5 : Répartition de l'utilisation des stratégies de traduction pour le système de TA Google Translate

Pour le troisième et dernier système de traduction, Systran, l'écriture inclusive et notamment les doubles formes abrégées ont posé de nombreux problèmes qui rendent les traductions parfois incompréhensibles.

Dans le Tableau 7, nous pouvons voir quelles ont été les stratégies de traduction les plus utilisées par le système de TA Systran selon la stratégie discursive.

	Doubles formes intégrales	Double formes abrégées	Règle de proximité	Créations langagières	Épicènes et formulations englobantes
Remplacement	33	123	1	11	0
Omission	0	13	4	2	0
Maintien	20	1	0	1	40

Tableau 7 : Stratégie de traduction selon la stratégie discursive pour le système de TA Systran.

Les doubles formes abrégées sont problématiques pour le système de traduction automatique notamment en raison de leur structure composite (avec point médian dans la plupart des cas). Bien que les doubles formes abrégées soient globalement remplacées par des formes neutres (telles que des épïcènes), on retrouve des restes de la structure française. Dans la traduction obtenue, on retrouve alors des points médians placés de manière un peu aléatoire, souvent autour de la traduction de la double forme abrégée comme avec « salarié-es » qui devient « ·wage earners » :

Exemple 23 :

Texte 8 : «*Nous nous souvenons de ses ordonnances pour casser le droit du travail et abaisser la protection des **saliarié-es**.* »

Traduction Systran : « We remember his orders to break the labor law and lower the protection of · **wage earners**. »

En plus des points médians, on a également des restes des morphèmes utilisées par les doubles formes abrégées. Ces morphèmes éparpillés autour des traductions empêchent non seulement la bonne compréhension du texte, mais faussent également la traduction des doubles formes abrégées. Dans certains cas, les morphèmes sont eux-mêmes traduits comme des formes à part entière comme par exemple « journalistes indépendant·e·s » qui est traduit par « independent journalists · the · » :

Exemple 24 :

Texte 3 : « *Nous, journalistes de terrain, **journalistes indépendant·e·s**, en poste en rédaction (...)* »

Traduction Systran : « *We, journalists on the ground, **independent journalists** · **the** · , working in the newsroom (...)* »

Dans d'autres cas encore, la double forme abrégée n'est même pas traduite et on retrouve seulement les points médians dans le texte traduit comme par exemple « effaré·e·s » qui est traduit par « · in · » :

Exemple 25 :

Texte 5 : « *Nous aussi, nous avons assisté, **effaré·e·s**, à des rites initiatiques réservés aux seuls jugés capables d'honorer la réputation d'associations imbibées de valeurs virilistes. »*

Traduction Systran : « *We too have witnessed, • **in** • , initiation rituals reserved only for those deemed capable of honoring the reputation of associations imbibed by virilistic values. »*

On peut alors dire dans ce cas-là qu'il y a eu une omission de la double forme abrégée. Un autre point à relever concerne le traitement de « professionnel·le·s » qui va être correctement traduit et remplacé par le neutre pluriel anglais « professionals » :

Exemple 26 :

Texte 6 : « *La confidentialité protège les femmes; inviter les **professionnel.le.s** à l'enfreindre est susceptible de les mettre encore plus en danger »*

Traduction Systran : « *Confidentiality protects women; inviting **professionals** to break it is likely to put them in even greater danger »*

Or dans d'autres cas, la même forme est maintenue avec la structure d'une double forme abrégée, mais traduite (« professional·e·s ») :

Exemple 27 :

Texte 6 : « *Beaucoup de femmes (mineures ou adultes) hésitent déjà à consulter parce qu'elles sont très mal reçues par **certain.e.s professionnel.le.s**. »*

Traduction Systran : « *Many women (minors or adults) already hesitate to consult because they are very poorly received by **some professional.e.s**. »*

Le système de traduction automatique manque ici donc à nouveau de cohérence, traduisant une même forme de différentes façons.

Concernant les doubles formes intégrales, différentes stratégies ont été utilisées. Le plus souvent, le système de traduction automatique remplace les doubles formes intégrales par des formes neutres (telles que des épécènes) déjà disponibles en anglais comme « un ou une » traduit par « a » :

Exemple 28 :

Texte 1 : « *Question pratique : où consulter **un ou une** psy ? »*

Traduction Systran : « *Practical question: where to consult **a** therapist? »*

Il arrive cependant que certaines doubles formes intégrales soient maintenues, cependant ces maintiens posent parfois problème, car on risque d'obtenir alors une répétition due aux formes neutres en anglais comme par exemple « mon/ma » traduit par « my/my » :

Exemple 29 :

Texte 1 : « J'ai également le droit d'interroger **mon/ma future** psy sur sa formation et sa méthode. »

Traduction Systran : « I also have the right to ask **my/my future** therapist about **his/her** training and method. »

Dans d'autres cas encore, uniquement la forme féminine va être maintenue comme avec « Français et Française » traduit par « French women » :

Exemple 30 :

Texte 7 : « On apprend dans ce sondage que 42 % des **Français et Françaises interrogées** ont, plus qu'avant, passé du temps à l'apéro. »

Traduction Systran : « This survey shows that 42% of **French women surveyed** spent more time on drinks than before. »

Les exemples d'utilisation de la règle de proximité sont tous omis et les créations langagières sont remplacées par des formes neutres (tels que des épïcènes). Enfin les épïcènes et les formulations englobantes sont maintenues comme l'anglais possède des équivalents neutres. Il arrive cependant que lorsque l'on a des formes neutres en français telles que « l' » ou « sa », l'anglais va remplacer cette forme par un masculin tel que « his » ce qui va alors à l'encontre de tout le processus de l'écriture inclusive :

Exemple 31 :

Texte 1 : « Vous avez aussi la liberté de l'interroger : sur **ses** tarifs, bien sûr, mais aussi sur **sa** formation, **sa** spécialisation, **sa** méthode. »

Traduction Systran : « You also have the freedom to question **him**: of course, but also on **his** training, **his** specialization, **his** method. »

Nous retrouvons ci-dessous (Figure 6) une représentation de la répartition de l'utilisation des stratégies de traduction (omission, remplacement, maintien, ajout) au sein de notre corpus traduit par Systran.

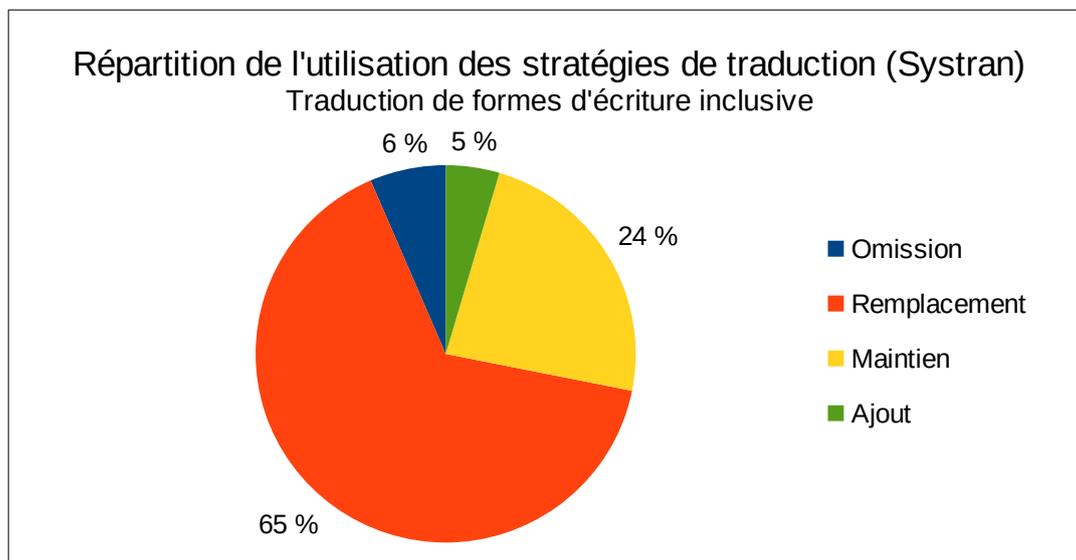


Figure 6 : Répartition de l'utilisation des stratégies de traduction pour le système de TA Systran

Si nous généralisons (Figure 7) les comportements des systèmes de TA et les stratégies de traduction utilisées, nous pouvons relever que la stratégie de traduction la plus utilisée est le remplacement. Cette observation s'explique par le fait que l'anglais, étant une langue grammaticalement non-genrée (contrairement au français), possède déjà des formes neutres ou épïcènes. Ces formes disponibles en anglais sont déjà inclusives voir même plus inclusives que certaines formes créées par l'écriture inclusive, car elles permettent d'aller au-delà de la binarité du genre. L'anglais va donc généralement remplacer toutes les formes françaises d'écriture inclusive. Bien que l'écriture inclusive soit remplacée, on ne peut dire que cela affecte véritablement le sens final du texte. L'inclusivité reste présente, néanmoins, on peut noter la suppression de l'effort de féminisation qui est réalisé lors de l'utilisation de l'écriture inclusive.

Une autre stratégie de traduction se démarque malgré tout et c'est le maintien. Le maintien de certaines formes d'écriture inclusive s'explique en partie par le maintien des épïcènes et formulations englobantes. Cette stratégie discursive qui est déjà généralement utilisée par l'anglais, permet des équivalences directes entre les épïcènes français et les formes neutres anglaises. Le maintien permet ainsi d'observer quels sont les termes qui sont déjà disponibles en anglais, notamment face aux créations langagières de l'écriture inclusive française.

Quant à l'omission et l'ajout, ces deux stratégies sont utilisées, mais elles font référence à des cas particuliers. L'omission est généralement utilisée face aux exemples de la règle de proximité comme l'accord de genre n'existe pas en anglais. L'ajout est utilisé afin de garder une certaine cohérence en anglais ou bien simplement afin de réintroduire les deux genres quand cela est nécessaire.

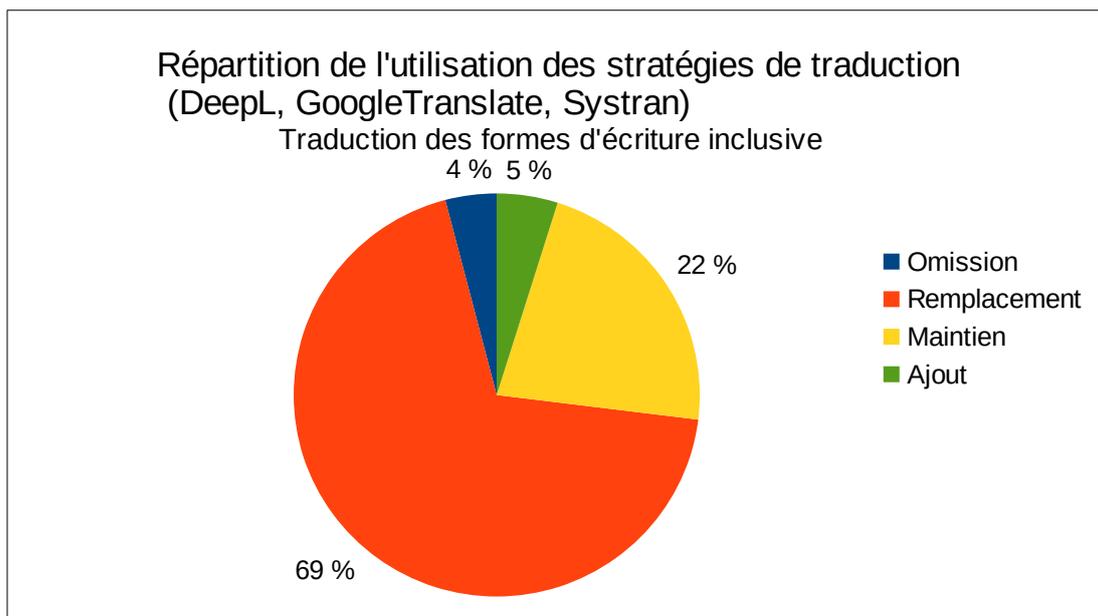


Figure 7 : Répartition de l'utilisation des stratégies de traduction pour les trois systèmes de TA

Enfin, si nous généralisons à présent les défis que pose l'écriture inclusive aux systèmes de TA, nous pouvons dire que les doubles formes abrégées et les doubles formes intégrales sont les deux stratégies les plus difficiles à traiter pour les trois systèmes de traduction automatique neuronaux. Les doubles formes abrégées posent d'abord problème en raison de leur forme composite complexe comprenant un symbole (dans la plupart des cas, le point médian). Ce symbole a notamment posé problème dans la bonne analyse des formes avec des erreurs de traduction liées au lexique. Les morphèmes accompagnant les doubles formes abrégées sont parfois traités comme des formes à part entière ce qui pose des problèmes de compréhension.

Les doubles formes intégrales posent de nombreux problèmes lorsqu'elles sont maintenues, car on retrouve alors deux formes identiques (deux formes neutres en anglais) dans la traduction. Parfois, il arrive que ces deux formes soient différentes, mais elles ne sont alors pas des traductions correctes ou pertinentes.

Ces deux stratégies discursives d'écriture inclusive peuvent à grande échelle fausser la traduction, voire même changer le sens d'une phrase ou d'un texte. Néanmoins, d'autres stratégies discursives comme la règle de proximité ou les épïcènes et les formulations englobantes, sont mieux traités par les systèmes de TA. Généralement, les exemples d'utilisation de la règle de proximité sont omis dans la traduction car l'anglais est une langue grammaticalement non-genrée. Les épïcènes et les formulations englobantes sont directement maintenues et ne posent alors aucun problème. Enfin, les créations langagières sont globalement bien identifiées par les systèmes de TA qui les remplacent par des formes neutres disponibles en anglais.

On pourrait en conclure que ce n'est pas directement l'écriture inclusive qui va poser problème aux systèmes de TA, mais bien certaines stratégies. Bien que nous ayons vu les différences de traitement entre les trois systèmes de traduction automatique, on retrouve partout des difficultés posées par les doubles formes abrégées et les doubles formes intégrales. On peut pourtant dire que le système de TA DeepL est celui qui est le moins affecté par l'écriture inclusive, ne connaissant presque aucun problème de traitement. Néanmoins c'est également le système de TA qui remplace le plus les exemples d'écriture inclusive. Ainsi d'un point de vue de la traduction, c'est celui qui est le plus efficace mais d'un point de vue de l'inclusivité le système de TA Google Translate semble meilleur notamment vis-à-vis des problèmes de traitements rencontrés par le système de TA Systran.

4.2.2. Analyse de l'écriture inclusive dans les traductions anglaises

Les différentes stratégies de traduction adoptées par les systèmes de traduction automatique vont avoir un rôle sur l'éventuelle présence de formes d'écriture inclusive en anglais. Nous avons repris les traductions simples de notre corpus obtenues après utilisation des trois systèmes de TA et nous nous sommes cette fois-ci intéressée à la façon dont l'écriture inclusive pouvait encore être présente en anglais.

Afin d'obtenir les formes d'écriture inclusive en anglais, nous n'avons pas essayé d'appliquer notre classification des stratégies discursives vue précédemment car celle-ci est uniquement applicable au français. Bien que certaines stratégies comme les doubles formes intégrales et les créations langagières existent également en anglais, d'autres stratégies discursives comme la règle de proximité ou les doubles formes abrégées n'existent pas car l'anglais est une langue non-genrée grammaticalement. Enfin les épiciènes et formulations englobantes sont la norme en anglais permettant une grande inclusivité (notamment au-delà de la binarité du genre) mais surtout de faire référence à des entités humaines de manière générale.

Pour obtenir le nombre de formes d'écriture inclusive demeurant après traduction simple (donc dans nos traductions anglaises), nous avons créé une nouvelle classification adaptée à la langue anglaise à partir des stratégies discursives disponibles en anglais (voir section 1.5). Afin d'avoir les véritables efforts conscients d'inclusivité et comme l'anglais est une langue genrée pronominalement, nous nous sommes intéressée uniquement aux pronoms à travers l'utilisation des doubles formes intégrales et des pronoms neutres au singulier (uniquement le singulier car le pluriel est le même peu importe le genre). Nous avons donc comme classification d'écriture inclusive en anglais :

- **Utilisation de pronoms neutres pour le singulier** : « they », « them », « their », « themself »
- **Doubles formes intégrales** : « she/he », « her/his »

A l'aide de cette classification, nous avons annoté, avec Glozz, les formes inclusives dans nos traductions obtenues.

Pour le système de TA DeepL, la présence de l'écriture inclusive passe principalement par les doubles formes intégrales avec 33 exemples au sein de notre corpus. Les pronoms neutres singulier ne sont utilisés que 2 fois dans le Texte 6. Bien que l'utilisation de cette forme pronominale neutre au singulier s'est grandement répandue ces dernières années, les doubles formes intégrales sont encore généralement préférées lorsque le genre est inconnu ou non pertinent. Nous avons ainsi au total dans notre corpus de traductions anglaises, 35 exemples d'écriture inclusive. Ce nombre, relativement faible, s'explique par le fait que dans la plupart des cas, les formes d'écriture inclusive ont été remplacées par des formes neutres (épïcènes) en anglais comme la langue le demande. Les éventuelles doubles formes intégrales et les pronoms neutres singulier demeurent grâce à des maintiens mais ils sont peu nombreux. Pour autant cela ne signifie pas que ces traductions ne sont pas inclusives. L'anglais permet l'inclusivité et l'ajout conscient de formes d'écriture inclusive notamment des doubles formes intégrales permet de rendre ces textes encore plus inclusifs. Néanmoins, de manière précise, le système de TA DeepL a une tendance à la neutralisation (que nous avons déjà remarqué dans la partie précédente) et non au maintien ce qui explique les résultats obtenus.

Pour le système de TA Google Translate, nous retrouvons à nouveau principalement des doubles formes intégrales avec 20 exemples au sein de notre corpus. Nous retrouvons des pronoms neutres singulier au sein du texte 1 et du texte 6. Nous avons au total dans notre corpus de traductions anglaises, 28 exemples d'écriture inclusive. Ce nombre relativement faible, s'explique de la même manière que pour le système DeepL. Néanmoins, on peut noter tout de même que la quantité de formes inclusives est encore plus faible et cela peut être dû au remplacement beaucoup plus fréquent des doubles formes intégrales par ce système de TA. Au lieu de les maintenir, les doubles formes intégrales sont remplacées par des tournures englobantes ou bien remplacées par des pronoms masculins ce qui va alors à l'encontre de la féminisation et de l'inclusivité.

Pour le système de TA Systran, nous retrouvons cette fois-ci un grand nombre de pronoms neutres singuliers avec 21 exemples au sein de notre corpus. Ces formes se retrouvent dans le texte 1 qui utilise ce pronom au lieu de doubles formes intégrales mises en avant en français dans le texte d'origine. Ce choix est intéressant et permet à ce système de TA de se démarquer des deux autres en choisissant une stratégie discursive plus moderne et surtout plus inclusive (au-delà de la binarité du genre). On retrouve tout de même 7 exemples de doubles formes intégrales mais la faible utilisation de cette stratégie discursive peut s'expliquer par le remplacement beaucoup plus fréquent des doubles formes intégrales par ce système de TA. Au lieu de les maintenir, les doubles formes intégrales sont remplacées par des tournures englobantes ou bien remplacées par des pronoms masculins (comme vu dans la partie précédente) ce qui va alors à l'encontre de la féminisation et de l'inclusivité. Nous avons au total dans notre corpus de traductions

anglaises, 28 exemples d'écriture inclusive. Ce nombre relativement faible, s'explique de la même manière que pour le système DeepL et Google Translate. Le système de TA Systran favorise le maintien de certaines structures composites comme les doubles formes abrégées or nous n'avons pas retenu ces formes pour notre analyse car elles n'ont finalement aucun sens grammatical en anglais et s'apparentent plutôt à des transpositions erronées.

Selon notre classification de stratégies discursives d'écriture inclusive en anglais, on retrouve principalement des doubles formes intégrales au sein de notre corpus traduit. Les pronoms neutres singulier sont plus rares et semblent dépendre plutôt du système de TA, comme nous l'avons vu avec Systran. Les doubles formes intégrales sont généralement des maintiens directs des doubles formes intégrales du texte d'origine et permettent ainsi une plus grande inclusivité.

Néanmoins certaines traductions, peu importe le système de TA, se retrouvent sans aucune forme d'écriture inclusive (d'après notre classification). Ces traductions (texte 2, texte 3, texte 4, texte 8) ne sont pas pour autant non inclusives car l'anglais est une langue grammaticalement non-genrée et dispose d'une syntaxe généralement neutre. Ainsi dans ces traductions, on ne retrouve pas d'effort conscient mais cela peut être dû à la nature même du texte d'origine ou à sa forme d'origine comme par exemple le texte 2 qui possède principalement des épïcènes et formulations englobantes.

Le nombre de formes d'écriture inclusive baisse donc fortement après la traduction simple de notre corpus, mais cela est encore une fois principalement lié aux différences caractéristiques des deux langues (langue genrée et langue non genrée) et n'est pas entièrement à cause des systèmes de TA.

4.2.3. Stratégies discursives après traduction inversée

Après les premières traductions des textes en utilisant les trois systèmes de TA, nous nous sommes penchée sur la traduction inversée afin d'observer plus en détails les procédés utilisés par ces systèmes lors de la traduction de formes d'écriture inclusive. Afin de réaliser ces traductions inversées (traduction dans la langue source, ici le français, des traductions obtenues dans la langue cible, ici l'anglais), nous avons à nouveau utilisé les trois systèmes de TA présentés auparavant : DeepL, Google Translate et Systran. Ces traductions inversées ont été réalisées le 11 décembre 2020 sur les traductions obtenues de nos 8 textes d'origine. Nous avons décidé d'observer quelles étaient les stratégies discursives d'écriture inclusive qui demeuraient après la traduction inversée. Nous avons repris la classification en cinq stratégies discursives présentée auparavant (p. 24).

Dans un premier temps, avec le système de traduction automatique DeepL, parmi les stratégies discursives d'écriture inclusive disponibles, on retrouve principalement (voire uniquement) des épïcènes et formulations englobantes dans les traductions inversées. Au

total, sur les 40 exemples d'écriture inclusive présents dans notre corpus après traduction inversée avec ce système de TA, 39 exemples correspondent à des épïcènes et formulations englobantes et 1 exemple correspond à une double forme abrégée.

Les autres stratégies discursives d'écriture inclusive ne sont donc plus présentes. On peut se demander si le maintien des épïcènes et formulations englobantes a une explication et si ce maintien est influencé par les stratégies discursives d'écriture inclusive qui demeurent après la première traduction avec le système de traduction automatique DeepL. On peut supposer que les épïcènes et les formulations englobantes sont aujourd'hui les formes d'écriture inclusive les plus courantes et les plus faciles d'utilisation par le grand public. Elles sont relativement communes même dans des textes sans volonté d'inclusivité et ainsi elles sont plutôt une spécificité de la langue française.

Si on s'intéresse de plus près aux formes d'écriture inclusive qui étaient encore présentes après la première traduction, on notera, comme nous l'avons vu auparavant, que le remplacement (qui neutralise généralement grâce à des épïcènes et des formulations englobantes) a été la stratégie de traduction la plus utilisée pour le système de TA DeepL. De plus, pour les éventuels épïcènes et formulations englobantes dans le texte d'origine, le maintien était la stratégie discursive la plus utilisée comme avec « personne » :

Exemple 32 :

Texte 1 : « *Justine vous donne tous ses conseils pour trouver **la personne** la plus apte à vous aider !* »

Traduction texte 1 : « *Justine gives you all her advice to find **the person** best suited to help you!* »

Traduction inversée texte 1 : « *Justine vous donne tous ses conseils pour trouver **la personne** la plus apte à vous aider !* »

Ainsi, de manière générale après la traduction du texte français en anglais, les épïcènes et les formulations englobantes représentaient déjà la majorité des formes restantes d'écriture inclusive.

Pour les autres stratégies discursives d'écriture inclusive qui auraient pu être maintenues en anglais, on retrouve globalement une omission ou un remplacement de ces formes-là comme la double forme intégrales « he or she » remplacée par « il » :

Exemple 33 :

Texte 5 : « *S'il ou **elle** le désire, l'accompagner dans sa démarche, juridiquement et psychologiquement.* »

Traduction texte 5 : « *If **he or she** wishes to do so, accompany **him or her** in **his or her** approach, legally and psychologically.* »

Traduction inversée texte 5 : « *S'il le souhaite, accompagnez-le dans sa démarche, juridiquement et psychologiquement.* »

La seule l'exception est l'exemple de double forme abrégée, cependant il semble dus au maintien du français dans la traduction « professionnels-les », ce qui a permis de transposer directement la forme lors de la traduction inversée :

Exemple 34:

Texte 3 : « (...) nous demandons à la Commission de la carte d'identité des journalistes **professionnels-les** (CCIJP) de tenir compte de l'évolution de nos métiers. »

Traduction texte 3 : « (...) we ask the Commission for the Identity Card of Professional Journalists - CCIJP (Commission de la carte d'identité des journalistes **professionnels-les** (CCIJP)) to take into account the evolution of our professions. »

Traduction inversée texte 3 : « (...) nous demandons à la Commission de la carte d'identité des journalistes **professionnels-les** (CCIJP) de prendre en compte l'évolution de nos métiers. »

On peut en conclure que le système de traduction automatique DeepL privilégie la neutralisation du texte et des formes d'écriture inclusive.

Dans un deuxième temps, avec le système de traduction automatique Google Translate, parmi les stratégies discursives d'écriture inclusive disponibles, on retrouve principalement des épïcènes et formulations englobantes dans les traductions inversées. Au total, sur les 39 exemples d'écriture inclusive présents dans notre corpus après traduction inversée avec ce système de TA, 34 exemples correspondent à des épïcènes et formulations englobantes, 3 exemples correspondent à des doubles formes intégrales, 1 exemple correspond à une double forme abrégée et un 1 exemple correspond à une création langagière. La plus grande quantité d'épïcènes et de formulations englobantes s'explique de la même façon que pour le système de traduction automatique DeepL, c'est-à-dire par le statut de ces formes et les stratégies de traduction utilisées par le système de traduction. Néanmoins, ici, on retrouve tout de même de légères variations, et pour les comprendre, on peut observer plus en détails nos résultats de traduction simple avec le système de traduction automatique Google Translate. Bien que la stratégie de traduction la plus utilisée est également le remplacement, on peut observer que certaines doubles formes intégrales ont été maintenues après la première traduction. Ainsi, cela peut expliquer les doubles formes intégrales présentes après la traduction inversée :

Exemple 35 :

Texte 1 : « Si vous entretenez une relation de confiance avec **lui/elle**, votre médecin généraliste peut également être de bon conseil et vous recommander **un ou une** thérapeute. »

Traduction texte 1 : « If you have a relationship of trust with **him / her**, your GP can also be of good advice and recommend a therapist. »

Traduction inversée texte 1 : « Si vous avez une relation de confiance avec **lui / elle**, votre médecin généraliste peut également vous conseiller et recommander un thérapeute. »

On pourrait également supposer que le système de traduction automatique Google Translate tend moins vers la neutralisation, mais plus vers un maintien de la forme d'origine. Ainsi si dans le texte source (ici notre texte traduit en anglais) des doubles formes intégrales ou des créations langagières étaient encore présentes, elles n'ont pas été transformées par le système.

Exemple 36 :

Texte 2 : « la **chercheuse** s'aperçoit que les équipes avec davantage de femmes réussissent mieux que celles composées majoritairement par des hommes. »

Traduction texte 2 : « (...) the **researcher** finds that teams with more women do better than those composed mainly of men. »

Traduction inversée texte 2 : « (...) la **chercheuse** constate que les équipes avec plus de femmes réussissent mieux que celles composées principalement d'hommes. »

On peut conclure, que le système de traduction automatique Google Translate, privilégie la neutralisation des formes d'écriture inclusive, mais réalise aussi un maintien de certaines formes d'origine.

Dans un dernier temps, avec le système de traduction automatique Systran, parmi les stratégies discursives d'écriture inclusive disponibles, on retrouve plusieurs formes d'écriture inclusive différentes dans les traductions inversées. Au total, sur les 43 exemples d'écriture inclusive présents dans notre corpus après traduction inversée avec ce système de TA, 28 exemples correspondent à des épïcènes et formulations englobantes, 8 exemples correspondent à des doubles formes abrégées, 5 exemples correspondent à des doubles formes intégrales, et 2 exemples correspondent à des créations langagières. En ce qui concerne les épïcènes et les formulations englobantes, la même réflexion que pour les deux autres systèmes de traduction automatique peut être appliquée. Néanmoins, la diversité des formes d'écriture inclusive demande une nouvelle analyse des résultats obtenus après la première traduction avec le système de TA Systran. Bien que le remplacement a été à nouveau la stratégie de traduction la plus utilisée, suivi par le maintien, il faut néanmoins rappeler que nous avons relevé un traitement particulier des doubles formes abrégées. Le point médian et les morphèmes avaient souvent posé problème au système de traduction automatique ce qui résulta en un plus grand nombre de maintiens de ces mêmes doubles formes abrégées. Cela est directement relié à une des particularités des traductions inversées obtenues avec le système de traduction automatique Systran qui est que, pour certains textes, on retrouve des restes de doubles formes abrégées avec des points médians dans le texte. Certains de ces points médians sont bien placés et permettent alors de retrouver des exemples de doubles formes abrégées :

Exemple 37 :

Texte 5 : « *Nous aussi, nous avons été jugé·e·s , catégorisé·e·s et stigmatisé·e·s .* »

Traduction texte 5 : « *We too have been judged · e · s, categorized · e · s and stigmatized · e · s.* »

Traduction inversée texte 5 : « *Nous aussi avons été jugés · e · s, catégorisés · e · s et stigmatisés · e · s.* »

Néanmoins, pour la majorité des cas, ces points médians sont placés de manière quasi-aléatoire empêchant la bonne compréhension du texte et créant même parfois la confusion :

Exemple 38 :

Texte 5 : « *Comme les étudiant·e·s actuel·le·s , (...).* »

Traduction texte 5 : « *As the students · he · of the present · the · , (...).* »

Traduction inversée texte 5 : « *En tant qu'étudiants · les · du présent · les · , (...).* »

Les problèmes liés aux points médians mal placés résultent directement du mauvais traitement de ces mêmes symboles lors de la première traduction.

Enfin, on retrouve dans la traduction inversée quelques exemples de doubles formes intégrales qui en plus d'avoir été maintenues lors de la première traduction, n'ont pas été supprimées ou neutralisées par la traduction inversée :

Exemple 39 :

Texte 1 : « *Il/elle pourra ainsi vous expliquer un peu plus longuement les raisons de ses choix.* »

Traduction texte 1 : « *He/she will be able to explain the reasons for his/her choices a little longer.* »

Traduction inversée texte 1 : « *Il/elle pourra expliquer un peu plus longtemps les raisons de ses choix.* »

On peut conclure que le système de traduction automatique Systran tend plus vers une neutralisation ou un maintien des formes d'écriture inclusive bien que cela peut se faire aux dépens de la qualité du texte.

D'une manière générale, sans se concentrer sur un système de traduction automatique en particulier, on a pu remarquer que parmi les cinq stratégies discursives d'écriture inclusive, les formes d'épicènes et de formulations englobantes sont les plus fréquentes après la traduction inversée. Que ce soit grâce à leur statut même ou bien à la stratégie de traduction la plus utilisée (le remplacement), ces formes étaient déjà les plus fréquentes après une traduction simple du français vers l'anglais. Ceci explique alors leur plus grande

présence dans la traduction inversée. De plus, ces formes sont généralement maintenues par le système de TA car des équivalents existent dans la langue.

Les doubles formes abrégées et les doubles formes intégrales ne sont néanmoins pas complètement absentes notamment avec le système de traduction automatique Systran. Ces formes plus spécifiques à l'écriture inclusive ont souvent été remplacées ou omises par les systèmes de TA. Néanmoins là où elles ont été maintenues, on les retrouve également après la traduction inversée. Ces formes-là étant plus récentes et plus complexes, en raison de la nature composite des doubles formes abrégées ou aux variations morphémiques des doubles formes intégrales qui sont souvent mal analysées par les systèmes de traduction automatique, deviennent les points clés sur lesquels les systèmes de TA neuronaux devraient se concentrer afin d'aller vers plus d'inclusivité. Les créations langagières ont généralement été remplacées (neutralisation) lors de la première traduction. Ainsi, il est rare de les retrouver après la traduction inversée sauf si le cotexte permet au système de traduction automatique d'identifier le genre, mais cela reste très rare. Enfin les exemples d'utilisation de la règle de proximité ont tous été remplacés (ou omis) par les premières traductions en raison des spécificités de l'anglais qui est une langue non-genrée et ne demande ainsi pas d'accord en genre. On ne retrouve aucun exemple de cette stratégie discursive après la traduction inversée.

L'écriture inclusive tout le long de notre étude (après les traductions simples et après les traductions inversées) est présente. Néanmoins, les systèmes de TA, que ce soit à travers les stratégies de traductions utilisées ou les formes d'écriture inclusive (ré)introduites, ont un impact sur l'inclusivité des textes.

4.3. Impact des systèmes de traduction automatique sur l'inclusivité

4.3.1. Présence du genre grâce aux pronoms

Les formes pronominales sont révélatrices de l'inclusivité d'un texte. Afin de mesurer l'inclusivité et la distribution du genre dans les textes de notre corpus, nous nous sommes inspirée des travaux de Prates *et al.* (2018), et nous avons ainsi compté les formes pronominales. Nous avons d'abord essayé avec l'outil TreeTagger, néanmoins la tokenisation posait problème pour certaines formes composées ainsi que les doubles formes intégrales. En conséquence, les formes pronominales étaient mal annotées et leur repérage était compliqué. Nous avons finalement utilisé le logiciel d'analyse de corpus Antconc²⁹ afin de repérer les pronoms dans les textes. Nous avons utilisé ce logiciel-ci car il nous permet de créer notre propre liste de mots à repérer ce qui facilite le processus.

Nous nous sommes basée pour le français sur la liste de formes pronominales suivantes : il, elle, ils, elles, celui, ceux, celle, celles, iel, iels. Nous avons décidé de compter les créations langagières « iel » et « iels » car ce sont des pronoms qui apparaissent dans

²⁹ Logiciel d'analyse de corpus Antconc : <http://www.laurenceanthony.net/software/antconc/>

notre corpus et qui sont aujourd'hui utilisés dans certains contextes. Comme nous nous intéressons seulement à la présence du genre renvoyant à des entités humaines, nous avons dû dans certains cas analyser le contexte des pronoms repérés dans les textes afin de retenir les formes pronominales pertinentes, ce qui fut possible avec l'outil concordance du logiciel Antconc. Pour notre classification des pronoms, nous avons créé trois catégories :

- Féminin : elle, elles, celle, celles
- Masculin : ils, ils, celui, ceux
- Neutre : iel, iels

Lorsque le pronom renvoie à un référent animé genré, le pronom est classé dans 'Féminin' ou 'Masculin'. Lorsque le pronom est une création langagière (tel que « iel » et « iels ») alors le pronom est classé dans 'Neutre'. Si le genre du référent n'est pas explicite ou défini mais que la forme est genrée comme par exemple « ils », nous classerons le pronom dans 'Masculin'.

Dans les textes d'origine français de notre corpus, nous obtenons les résultats suivants (Figure 8) après avoir compté et classé les pronoms.

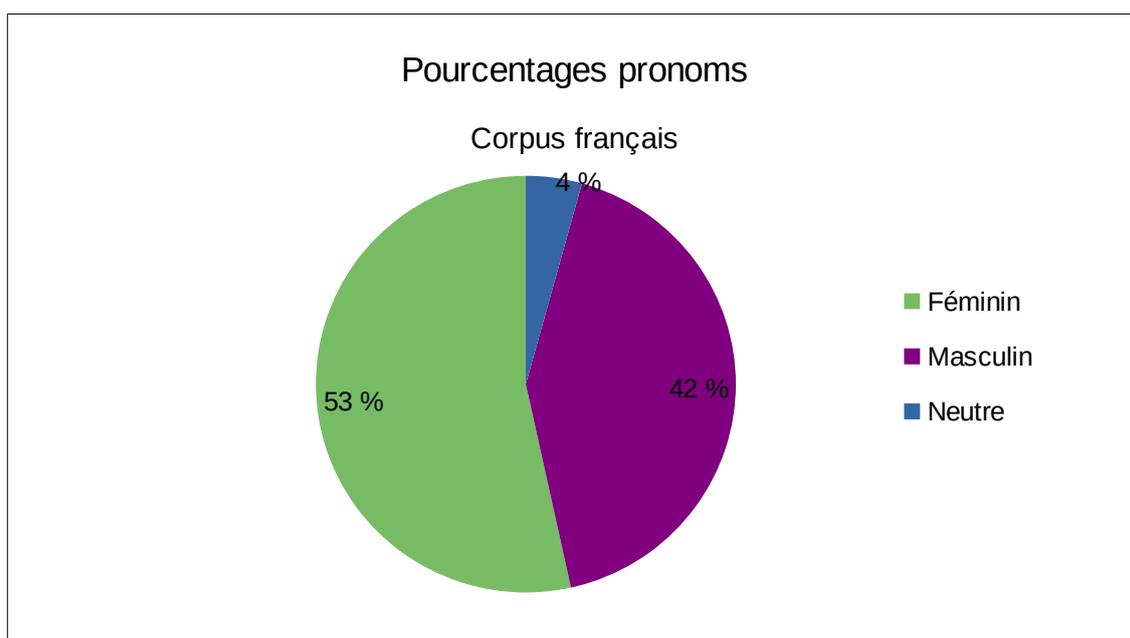


Figure 8 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus français

Au total, au sein du corpus français, on retrouve 118 formes pronominales correspondant à notre classification. On peut observer une division presque équitable entre le genre masculin et le genre féminin. Néanmoins, il y a tout de même un plus grand nombre de pronoms féminins au sein de notre corpus. Ce résultat est une bonne chose car il prouve que l'effort de féminisation qu'encourage l'écriture inclusive est présent même dans les formes pronominales. La part des pronoms neutres est très faible car les créations langagières neutres sont encore peu usitées dans le langage courant.

Afin de compter les formes pronominales dans les traductions anglaises obtenues après utilisation des systèmes de TA, nous nous sommes basée pour l'anglais sur la liste de formes pronominales suivantes : *she*, *he*, *they*, *the one*, *the ones*, *those*. Nous avons à nouveau repéré et compté les formes pronominales grâce au logiciel d'analyse de corpus Antconc.

Nous avons à nouveau créé une classification pour les pronoms avec trois catégories :

- Féminin : *she*
- Masculin : *he*
- Neutre : *they*, *the one*, *the ones*, *those*

La répartition des pronoms pour l'anglais est plus évidente. Le pronom neutre pluriel « *they* » (qui peut être utilisé aussi au singulier) a été classé dans 'Neutre', tout comme « *the one* », « *the ones* », et « *those* » lorsqu'ils renvoient à des référents animés. Les pronoms « *he* » et « *she* » ont été classés selon le genre auquel ils renvoient.

Pour le système de traduction automatique DeepL, nous obtenons les résultats suivants (Figure 9) après avoir compté et classé les pronoms.

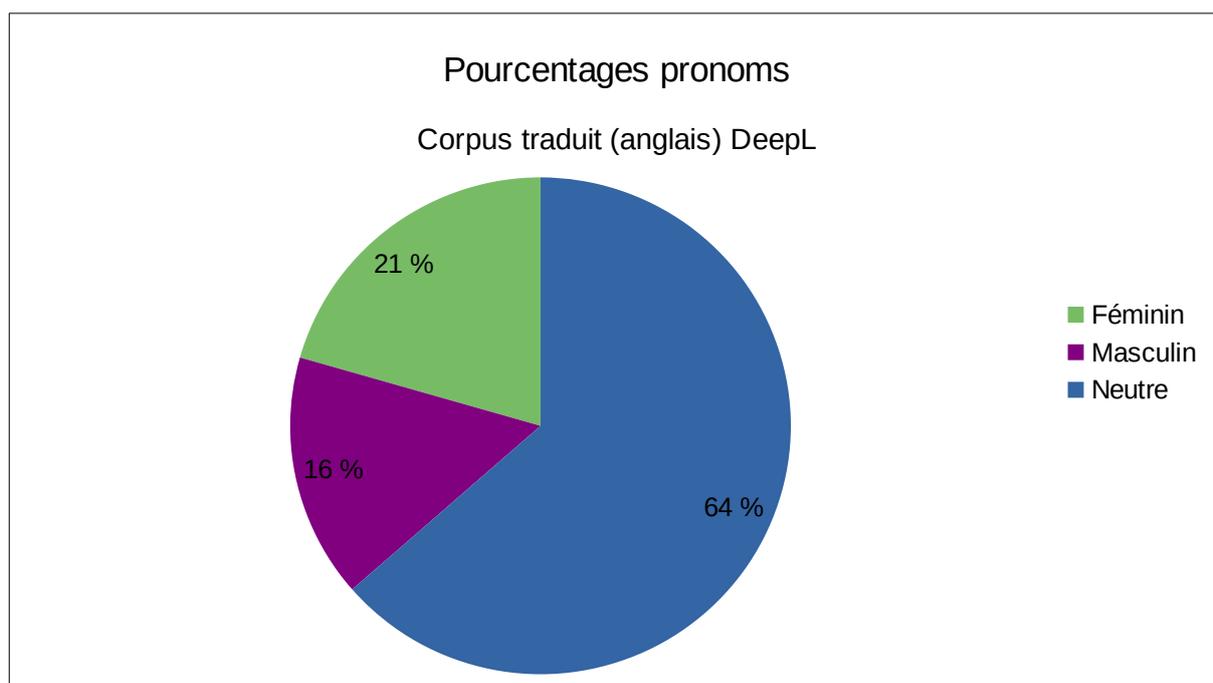


Figure 9 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus traduit avec le système de TA DeepL

Au total au sein du corpus, après traduction avec le système de TA DeepL, on retrouve 107 formes pronominales correspondant à notre classification. La catégorie 'Neutre' est la plus importante et cela peut s'expliquer par le fait que cette catégorie comprend aussi bien les formes singulier neutre (« *they* » utilisé pour désigné une seule personne ou « *the one* ») que les formes pluriel (neutre) exprimés par les formes pronominales « *they* »,

« the ones », « those ». Pour les catégories genrées binairement ('Féminin' et 'Masculin'), le féminin est plus important. Les textes du corpus, traduits par le système de TA DeepL, semblent plutôt bien équilibrés et surtout inclusifs.

Pour le système de traduction automatique Google Translate, nous obtenons les résultats suivants (Figure 10) après avoir compté et classé les pronoms.

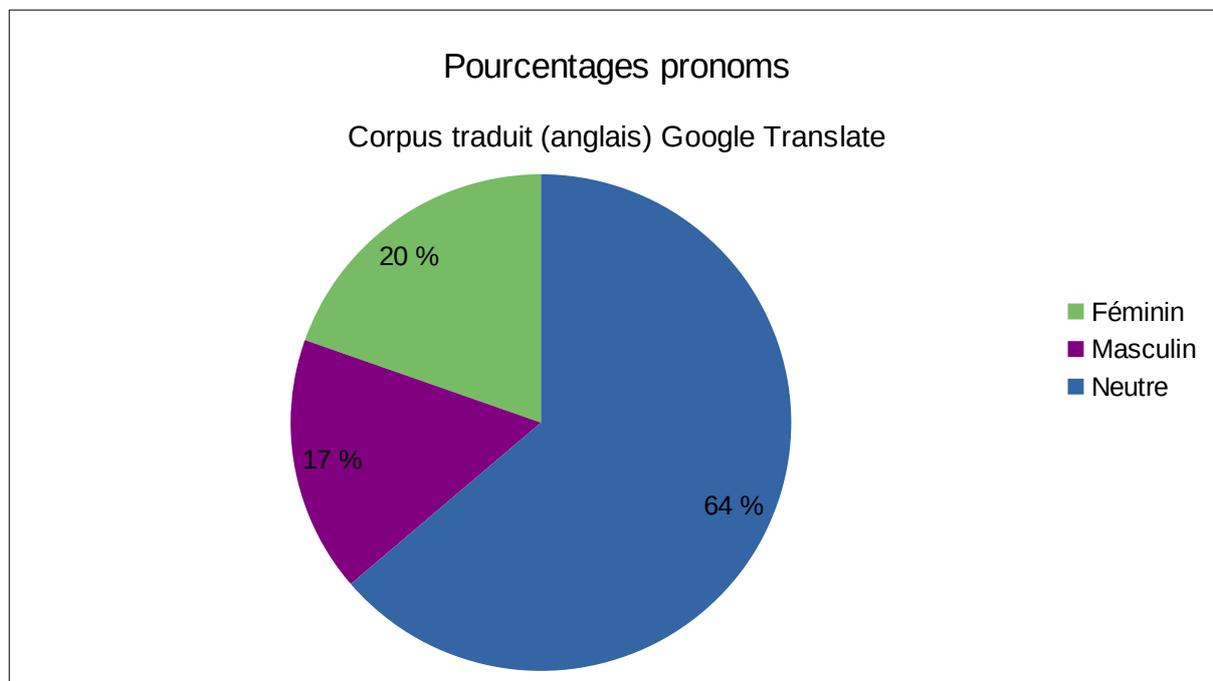


Figure 10 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus traduit avec le système de TA Google Translate

Au total au sein du corpus, après traduction avec le système de TA Google Translate, on retrouve 102 formes pronominales correspondant à notre classification. Les résultats sont similaires à ceux obtenus avec le système de TA DeepL, la catégorie 'Neutre' étant la plus importante, et la catégorie 'Féminin' étant un peu plus importante que 'Masculin' notamment grâce aux maintiens des formes pronominales féminines.

Pour le système de traduction automatique Systran, nous obtenons les résultats suivants (Figure 11) après avoir compté et classé les pronoms.

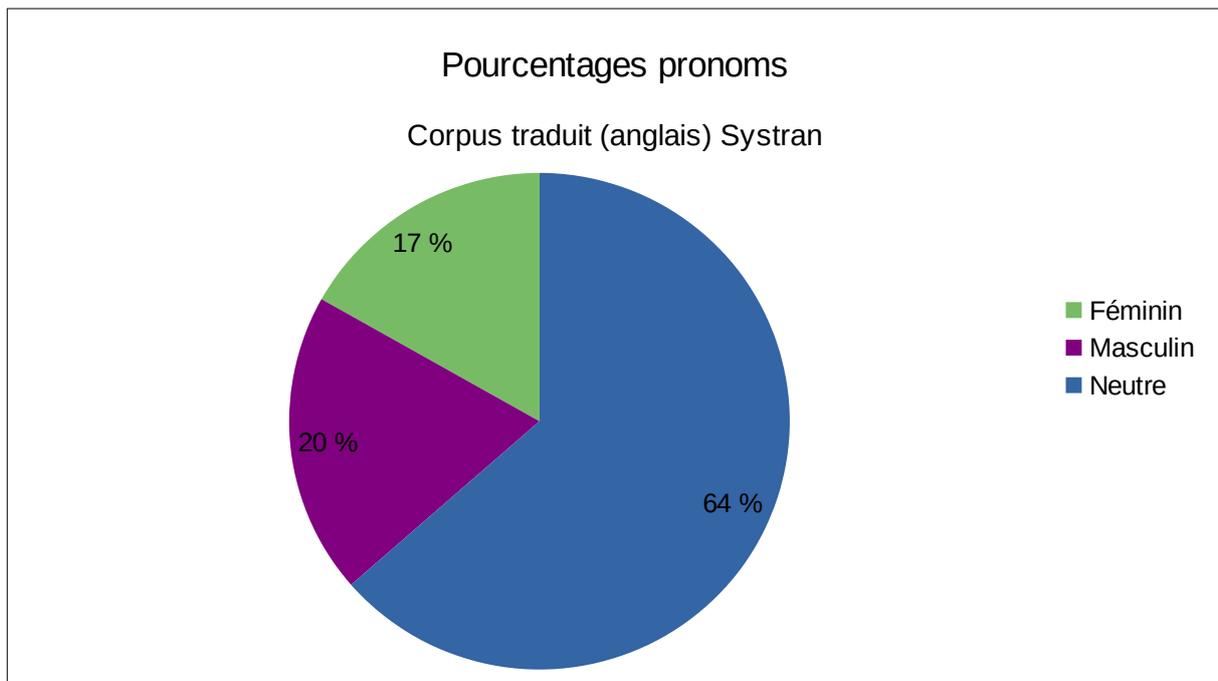


Figure 11 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus traduit avec le système de TA Systran

Au total au sein du corpus, après traduction avec le système de TA Systran, on retrouve 107 formes pronominales correspondant à notre classification. A nouveau la catégorie 'Neutre' est la plus importante ce qui est positif du point de vue de l'inclusivité notamment en dehors de la binarité du genre. Cependant, on remarque aussi que le pourcentage correspondant à la catégorie 'Masculin' est plus élevé que le pourcentage de la catégorie 'Féminin'. Les traductions réalisées par ce système semblent donc plus pencher vers le genre masculin.

Globalement, les traductions anglaises obtenues avec les systèmes de TA ont presque autant de formes pronominales que les textes d'origine. Les nombreux ajouts réalisés par certains systèmes de TA compensent les omissions et les reformulations des formes pronominales en anglais. La répartition des pronoms semble plus pencher vers le féminin sauf pour le système Systran. Cette masculinisation de textes inclusifs est négative car elle montre un mauvais traitement des formes inclusives de la part du système de TA. Le point positif, qui est néanmoins surtout relié à la langue cible (anglais), est l'augmentation de l'utilisation de formes neutres.

Enfin, nous avons repris la classification des pronoms pour les textes français afin d'observer leur répartition dans les textes obtenues après traduction inversée.

Pour le système de TA DeepL, après la traduction inversée, nous obtenons les résultats suivants (Figure 12) après avoir compté et classé les pronoms :

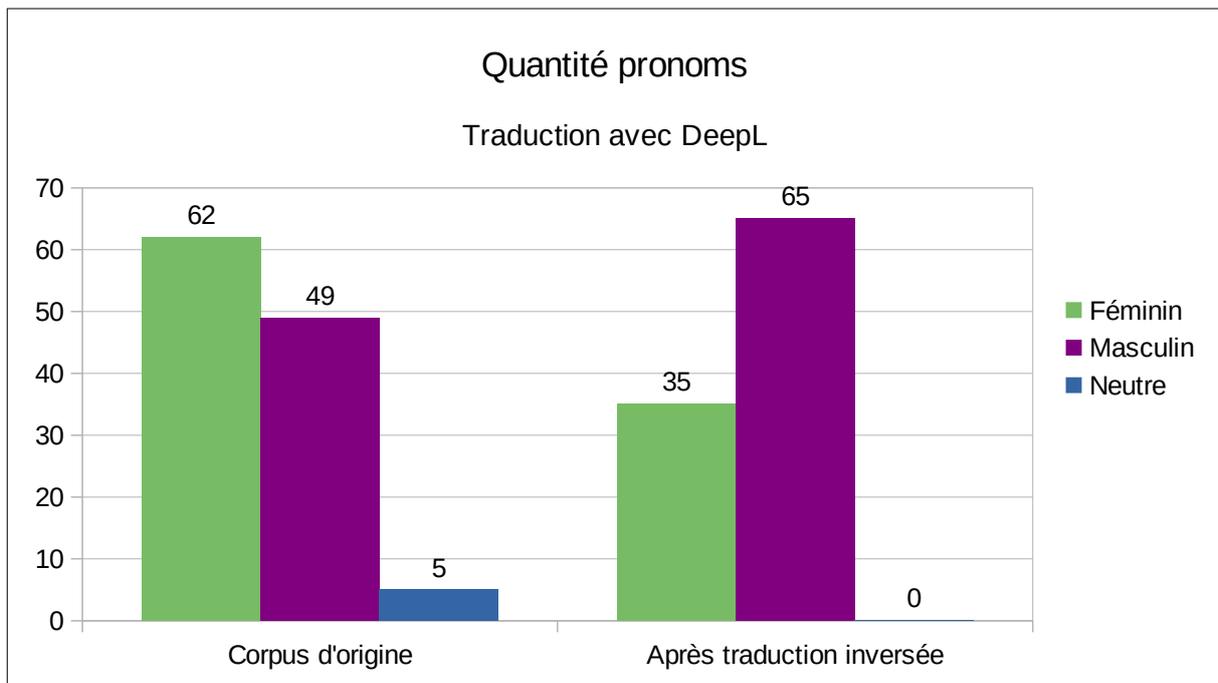


Figure 12 : Représentation du nombre de pronoms pour chaque catégorie dans le corpus d'origine et le corpus de traduction inversée avec le système de TA DeepL

Au total au sein du corpus, après traduction inversée avec le système de TA DeepL, on retrouve 100 formes pronominales correspondant à notre classification. La catégorie 'Neutre' ne comporte aucune forme pronominale ce qui montre que les créations langagières « iel » et « iels » qui avaient été traduites en anglais par des formes neutres, n'ont pas été retraduites comme telles avec le système de TA DeepL. Ces formes ont généralement été remplacées par des formes masculines. Le nombre le plus important correspond à la catégorie 'masculin'. Les traductions inversées semblent ainsi être plus masculines et moins inclusives notamment par rapport à la répartition des pronoms dans les textes d'origine.

Pour le système de TA Google Translate, après la traduction inversée, nous obtenons les résultats suivants (Figure 13) après avoir compté et classé les pronoms :

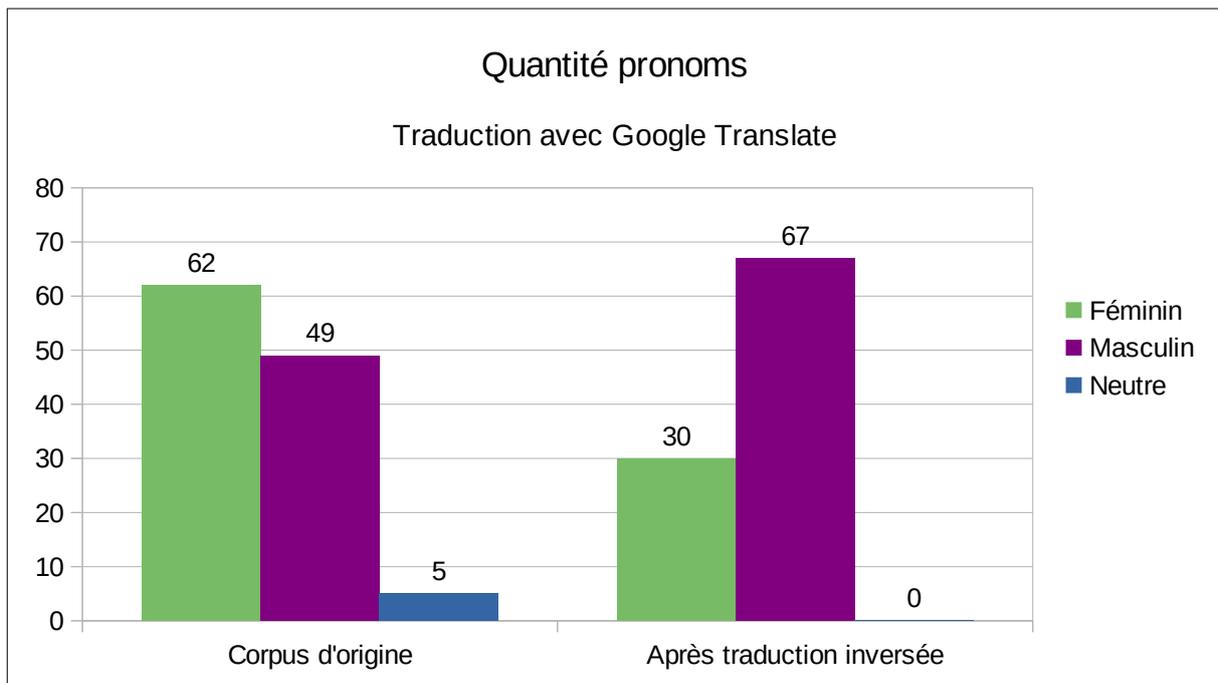


Figure 13 : Représentation du nombre de pronoms pour chaque catégorie dans le corpus d'origine et le corpus de traduction inversée avec le système de TA Google Translate

Au total au sein du corpus, après traduction inversée avec le système de TA Google Translate, on retrouve 97 formes pronominales correspondant à notre classification. A nouveau, la catégorie 'Neutre' ne comporte aucune forme pronominale ce qui montre que les créations langagières « iel » et « iels » qui avaient été traduites en anglais par des formes neutres, n'ont pas été retraduites comme telle avec le système de TA Google Translate. La part des pronoms appartenant à la catégorie 'Masculin' est à nouveau très élevée notamment face à la part 'Féminin'. Les textes obtenus après la traduction inversée réalisée par le système de TA Google Translate sont donc également plus masculins. Ce grand écart entre les deux catégories genrées est d'autant plus frappant en comparaison avec les résultats obtenus aux étapes précédentes. On peut supposer que la majorité des formes neutres anglaises telles que « they » ont été traduites par des formes masculines, ce qui expliquerait la hausse du pourcentage de pronoms classés dans 'Masculin'.

Pour le système de TA Systran, après la traduction inversée, nous obtenons les résultats suivants (Figure 14) après avoir compté et classé les pronoms :

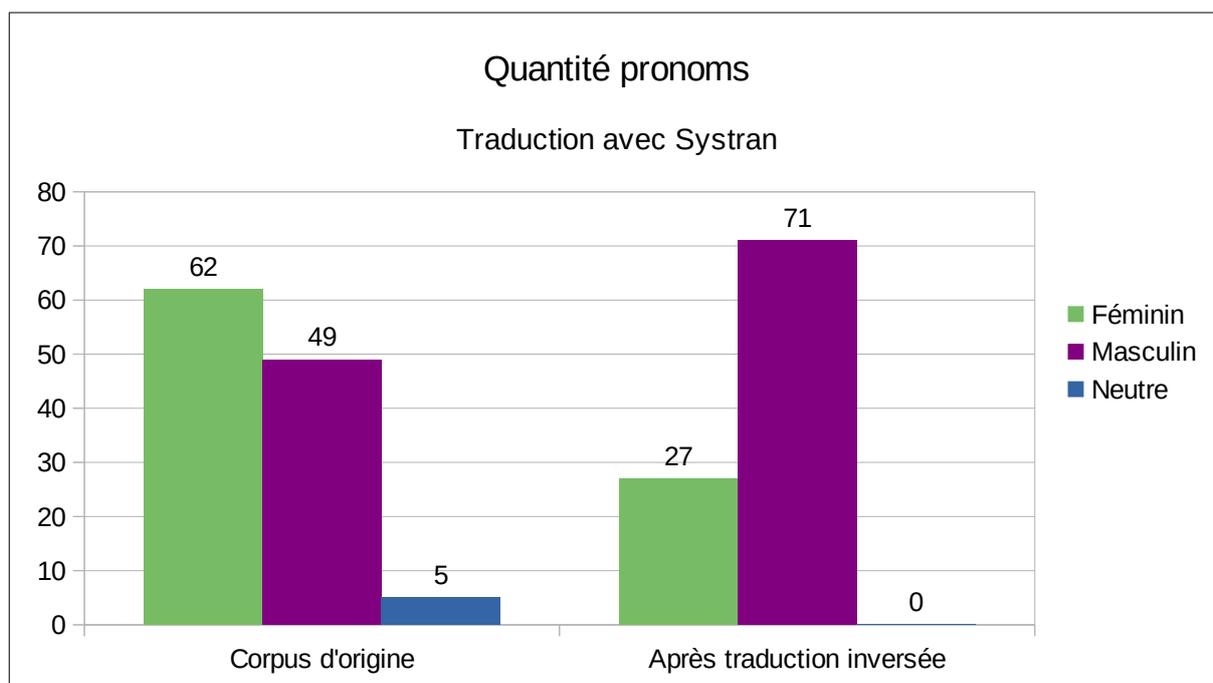


Figure 14 : Représentation du nombre de pronoms pour chaque catégorie dans le corpus d'origine et le corpus de traduction inversée avec le système de TA Systran

Au total au sein du corpus, après traduction inversée avec le système de TA Systran, on retrouve 98 formes pronominales correspondant à notre classification. A nouveau, la catégorie 'Neutre' ne comporte aucune forme pronominale ce qui montre que les créations langagières « iel » et « iels » qui avaient été traduites en anglais par des formes neutres, n'ont pas été retraduites comme telles avec le système de TA Systran. Le nombre de pronoms appartenant à la catégorie 'Masculin' est très élevé notamment par rapport au nombre de la catégorie 'Féminin' ce qui montre une forte masculinisation des textes après traduction inversée avec le système de TA Systran. Cette tendance à la masculinisation du système Systran se retrouve après la traduction simple avec un pourcentage plus élevé déjà, par rapport aux autres systèmes de TA.

Globalement, après la traduction inversée, on retrouve une diminution du nombre de formes pronominales peu importe le système de TA, une tendance qui était déjà observée après la traduction simple. Les formes pronominales neutres ont toutes disparu et ainsi celles qui étaient présentes dans les textes d'origine français, n'ont pas été maintenues après la traduction inversée. Ceci est d'autant plus visible car après les traductions simples, les textes anglais contenaient une majorité de pronoms neutres. Ces pronoms neutres singuliers et pluriels ont été globalement transformés en formes masculines ce qui explique les nombres élevés de pronoms appartenant à la catégorie 'Masculin'. Si on compare les trois systèmes de TA après traduction inversée, le système de TA Systran est celui qui a le nombre de pronoms appartenant à la catégorie 'Masculin' le plus élevée.

Ainsi, après les traductions inversées, les textes obtenus sont beaucoup moins inclusifs et au contraire beaucoup plus marqués en genre masculin au niveau des pronoms.

Finalement, nous pouvons à présent observer plus précisément l'impact des systèmes de TA sur les genres des pronoms à l'aide du graphique suivant (Figure 15).

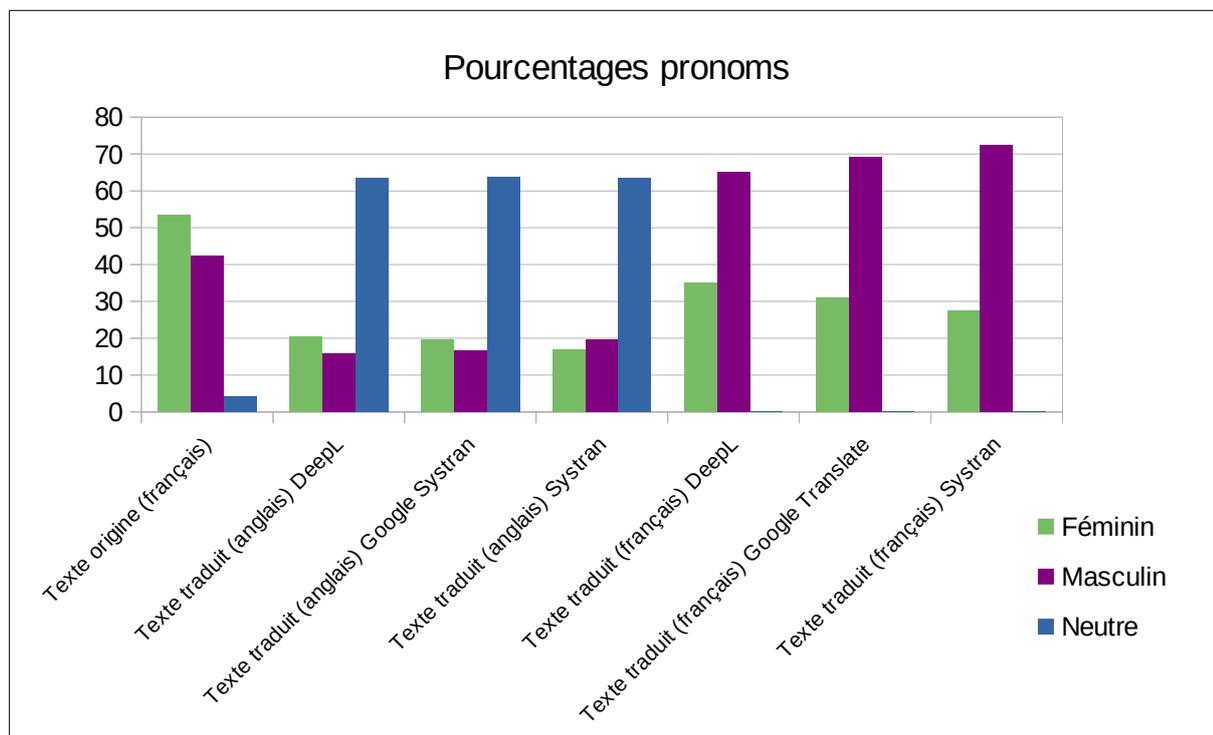


Figure 15 : Représentation du pourcentage de pronoms pour chaque catégorie dans le corpus aux différentes étapes de notre étude avec les trois systèmes de traduction automatique

L'un des éléments les plus frappants lorsqu'on observe la progression quantitative des formes pronominales est le renversement que l'on observe avec la catégorie 'Neutre'. Tandis qu'on avait très peu de formes pronominales neutres dans les textes d'origine, on en retrouve une grande quantité dans les traductions, or cela s'explique en partie par le fait que les formes plurielles anglaises ont été classées dans cette catégorie car ces formes pronominales ne renvoient à aucun genre et de ce fait sont vues comme neutre dans notre classification. Néanmoins, après la traduction inversée, toutes ces formes pronominales disparaissent au profit de formes masculines plurielles ou singulières. On passe par exemple d'une phrase inclusive en anglais, avec deux pronoms pluriels neutres « those » (ce qui garde l'inclusivité de la phrase d'origine composée de doubles formes intégrales), à une phrase masculinisée en français, avec deux pronoms pluriels masculins « ceux » ce qui réintroduit le genre (masculin) là où il n'était pas présent en anglais :

Exemple 40 :

Texte 5 : « Ce «nous» n'est pas homogène, mais il veut rassembler tou-te-s **celles** et **ceux** qui ont subi ces agressions, **celles** et **ceux** qui se sentent concerné-e-s par l'urgence d'un changement de culture dans ces écoles (...)»

Traduction Systran texte 5 : « This "we" is not homogeneous, but it wants to bring together all · and · **those** who have been subjected to these attacks, **those** who feel concerned · the urgency of a change of culture in these schools (...)»

Traduction inversée Systran : « Ce "nous" n'est pas homogène, mais il veut rassembler tous les · et · **ceux** qui ont subi ces attaques, **ceux** qui se sentent concernés · l'urgence d'un changement de culture dans ces écoles (...)»

La catégorie 'Masculin' est ainsi très importante après la traduction inversée. Cette inégalité dans la représentation genrée des pronoms est d'autant plus visible avec le système Systran après la traduction inversée. Enfin, on voit bien que les systèmes de TA vont avoir un impact important sur la présence de certaines formes pronominales car le léger déséquilibre entre la catégorie 'Masculin' et 'Féminin' qui était au profit du 'Féminin', va se retourner pour être grandement au profit du 'Masculin' après les traductions simples.

4.3.2. Progression quantitative de l'écriture inclusive

Afin de voir quel est l'impact des systèmes de traduction automatique non plus seulement sur les pronoms mais sur l'écriture inclusive, nous avons observé la progression quantitative des formes d'écriture inclusive. Nous avons ainsi relevé le nombre de formes d'écriture inclusive dans chaque texte avant toute traduction, après la traduction simple et enfin après traduction inversée, pour chaque système de traduction automatique.

Avant toute traduction, le nombre de formes d'écriture inclusive relevées dans notre corpus était de 249. Après la première traduction (en anglais) puis la traduction inversée (en français), on peut observer une importante diminution du nombre de formes d'écriture inclusive peu importe le système de traduction automatique comme le montre le Tableau 8 ci-dessous :

Système de traduction automatique	Nombre de formes d'écriture inclusive dans notre corpus d'origine	Nombre de formes d'écriture inclusive dans notre corpus après la traduction (en)	Nombre de formes d'écriture inclusive dans notre corpus après la traduction inversée (fr)
DeepL	249	35	40
Google Translate	249	28	38
Systran	249	28	43

Tableau 8 : Nombre de formes d'écriture inclusive avant et après traduction selon le système de TA

La diminution du nombre de formes d'écriture inclusive a été importante avec les trois systèmes de traduction automatique notamment après la traduction simple et cela est dû aux stratégies de traduction adoptées par le système. Ce qui est encore plus marquant c'est la grande diminution entre le corpus d'origine, puis le corpus traduit.

Cette forte diminution du nombre de formes d'écriture inclusive entre le corpus d'origine et le corpus transformé (traduction simple et traduction inversée) nous invite à observer plus en détails les textes de notre corpus indépendamment du système de traduction automatique. Le Tableau 9 ci-dessous permet de visualiser plus précisément la différence quantitative pour chaque texte du corpus avant et après les traductions :

Texte du corpus	Nombre de formes d'écriture inclusive dans notre corpus d'origine	Nombre de formes d'écriture inclusive dans notre corpus après la traduction (en)			Nombre de formes d'écriture inclusive dans notre corpus après la traduction inversée (fr)		
		DeepL	Google Translate	Systran	DeepL	Google Translate	Systran
Texte 1	37	19	21	21	3	4	6
Texte 2	19	0	0	0	10	12	13
Texte 3	19	0	0	0	1	1	0
Texte 4	3	0	0	0	0	0	1
Texte 5	100	3	2	3	8	4	5
Texte 6	53	12	5	4	14	14	14
Texte 7	5	1	0	0	1	1	2
Texte 8	13	0	0	0	3	3	2

Tableau 9 : Nombre de formes d'écriture inclusive avant et après traductions selon le système de TA et selon le texte

On remarque des comportements particuliers tel qu'un passage à 0 formes d'écriture inclusive après traduction simple pour certains textes comme les textes 2 et 8 avant de voir des nombres plus élevés après traduction inversée. Ceci peut s'expliquer par le fait que l'annotation des formes d'écriture inclusive en anglais est basé seulement sur deux stratégies discursives, et les épïcènes et formulations englobantes (qui sont des caractéristiques de l'anglais) n'ont pas été relevés car ils ne sont pas vus comme des effort conscient d'inclusivité en anglais.

Néanmoins, il est évident que la traduction simple est responsable d'une forte baisse du nombre des formes d'écriture inclusive ce qui aura ensuite une répercussion sur le nombre de formes restantes après traduction inversée.

Afin de voir si il y a une relation entre la stratégie discursive d'écriture inclusive utilisée dans un texte et le nombre de formes d'écriture inclusive présentes après les traductions, nous avons également représenté (Figure 16) le nombre de chaque stratégie discursive dans chaque texte du corpus d'origine. Nous avons fait de même (Figure 17, Figure 18, Figure 19) à partir des traductions inversées pour les trois systèmes de TA.

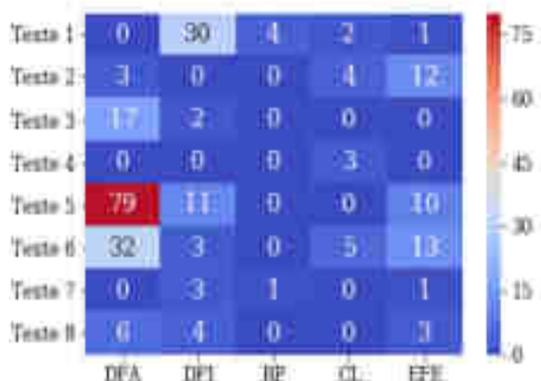


Figure 16 : Matrice représentant la quantité de formes de chaque stratégie discursive par texte dans le corpus d'origine

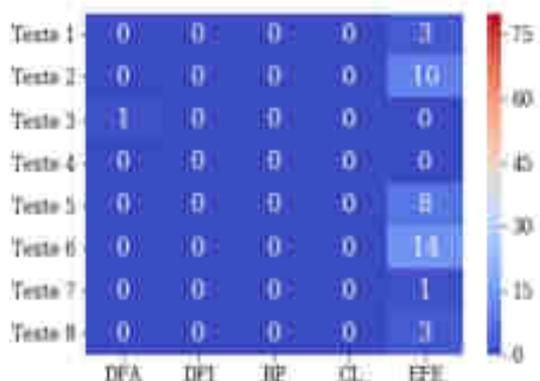


Figure 17 : Matrice représentant la quantité de chaque stratégie discursive par texte après traduction inversée avec DeepL

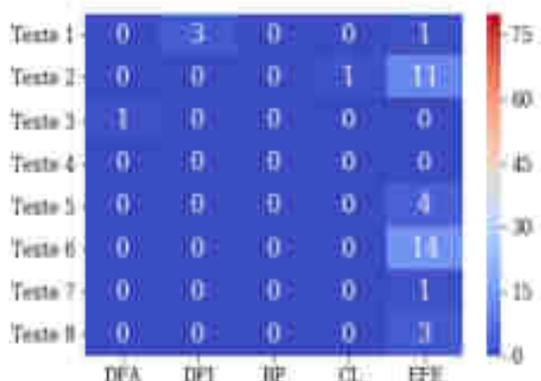


Figure 18 : Matrice représentant la quantité de chaque stratégie discursive par texte après traduction inversée avec Google Translate

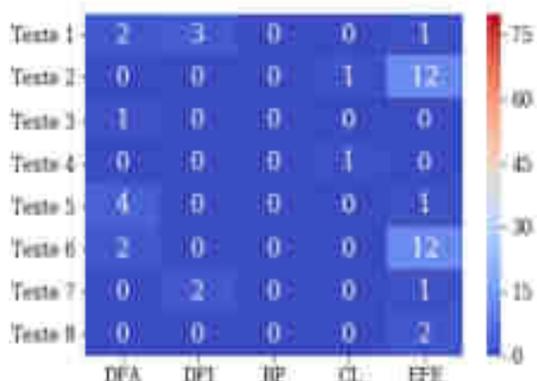


Figure 19 : Matrice représentant la quantité de chaque stratégie discursive par texte après traduction inversée avec Systran

Certains textes comme le texte 2 n'ont pas vu le nombre de formes d'écriture inclusive diminuer fortement. Ceci s'explique par la nature des formes d'écriture inclusive dans ce texte qui sont principalement des épécènes et formulations englobantes comme « individus », « membres » ou « collègues », des formes globalement maintenues. Au contraire, dans les textes 3 ou 5 on retrouve surtout des exemples de doubles formes abrégées ou de doubles formes intégrales, qui sont généralement remplacées (neutralisées) lors de la traduction automatique. Ces remplacements expliquent qu'après deux traductions, ces textes n'ont même pas la moitié de leur nombre de départ même si le système de TA Systran a maintenu quelques doubles formes abrégées. Ce système est celui qui a permis le maintien de quelques formes de cette stratégie discursive même si globalement sa présence a fortement diminué. Enfin, le texte 1 qui était riche en doubles

formes intégrales se retrouve avec très peu de formes d'écritures inclusives exceptés quelques maintiens avec les systèmes de TA GoogleTranslate et Systran.

4.4. Conclusion de l'étude

Pour conclure notre étude, l'écriture inclusive va être inévitablement altérée par les systèmes de traduction automatique. Cependant, comme nous l'avons vu, cela peut différer selon le système. L'écriture inclusive, sous toutes ses formes, va subir des transformations selon les stratégies de traduction adoptées par le système de TA. Certaines de ces stratégies réduisent la présence de l'inclusivité, la renforcent voire même la renversent. Cependant, l'écriture inclusive grâce aux particularités langagières de la langue cible (l'anglais) que nous avons choisi parvient à survivre même si en beaucoup plus faible quantité après deux passages dans un système de TA. L'écriture inclusive qui passe en grande partie par les formes pronominales en anglais, est affectée par des tendances à la masculinisation par certains systèmes de TA (Google Translate et Systran) notamment face à des formes neutres anglaises lors de la traduction inversée. Ces comportements vont à l'encontre de l'inclusivité et cela s'observe de manière générale peu importe les stratégies discursives. Après une traduction simple, il est évident que plus de la moitié des formes d'écriture inclusive ont été remplacées. Ces remplacements ne sont souvent pas au détriment de l'inclusivité car ils deviennent des formes neutres en anglais, qui pour beaucoup sont plus inclusifs que la féminisation porté par l'écriture inclusive, car le neutre peut aller au-delà de la binarité du genre. Le remplacement est finalement pour les trois systèmes de traduction automatique la stratégie de traduction la plus utilisée peu importe la stratégie discursive. Le maintien, qui apparaît comme la stratégie de traduction idéale, est principalement utilisé lorsqu'on a la possibilité dans la langue source d'avoir une adaptation grammaticale, notamment avec les épécènes et formulations englobantes, ou bien certaines doubles formes intégrales. L'écriture inclusive face à la traduction automatique pose ainsi indéniablement des problèmes, que ce soit au niveau de la présence et forme de l'écriture inclusive après traduction ou bien au niveau de la lisibilité et grammaticalité du texte qui peuvent être remises en question en raison de difficultés de traitement.

Les progrès des systèmes de traduction automatique réalisés jusqu'à présent nous prouvent malgré tout qu'il est possible de perfectionner les résultats obtenus. Enfin, on peut supposer que c'est par une plus grande utilisation de l'écriture inclusive et des différentes stratégies discursives par le grand public que les systèmes de traduction automatique vont pouvoir s'améliorer. Néanmoins, afin d'obtenir un texte bien traduit et inclusif, nous recommandons d'utiliser plutôt les doubles formes intégrales et les épécènes et formulations englobantes qui sont des stratégies discursives généralement maintenues et bien traitées par les systèmes de TA.

CONCLUSION

L'écriture inclusive (ou le langage inclusif) est à la fois source et conséquence de réflexions linguistiques. Les propositions de différentes conventions permettant une plus grande égalité dans la langue, ont été reçues de manière positive et négative dans divers domaines, mais c'est au domaine du traitement automatique des langues que nous nous sommes intéressée. Que ce soit les nouvelles formes (liés au point médian et aux doubles formes abrégés), une diversification du lexique (néologismes, mots épiciens, créations), un retour de la règle de proximité ou une plus large utilisation de la neutralisation, le TAL doit faire face à de nouveaux défis et s'adapter. Les différents outils permettant la tokenisation, la correction ou encore la traduction de texte français, commencent malgré tout à offrir des solutions aux difficultés langagières.

En observant de plus près un outil en particulier, la traduction automatique, nous avons pu voir que des défis liés à la question des genres existent déjà. Ces défis s'accompagnent d'autres difficultés que rencontrent les systèmes de traduction automatique face à l'écriture inclusive. Les stratégies utilisées par le système de traduction automatique répondent globalement aux différents types de stratégies discursives mais également aux particularités de la langue cible. Le remplacement des formes d'écriture inclusive est dans notre cas pas toujours préjudiciable à l'inclusivité mais déplace alors la volonté d'une féminisation de la langue vers une neutralisation. La multiplicité des stratégies discursives disponibles permet néanmoins de s'adapter aux éventuelles difficultés des outils de TAL et de contourner les problèmes actuels en attendant de meilleurs résultats. Parmi ces problèmes, on retrouve les biais de genre dont la réduction est une question actuelle dans le domaine du TAL, avec des travaux comme ceux de Sun *et al.* (2019) qui cherchent à comprendre d'où viennent ces biais et comment on peut les éviter (par manipulation du corpus, des algorithmes, ou d'autres paramètres). Les travaux de Saunders *et al.* (2020a; 2020b) qui se focalisent principalement sur la traduction automatique (avec comme langues sources et cibles : anglais, allemand, espagnol, hébreu), montrent que des solutions sont possibles (apprentissage par transfert (Saunders & Byrne, 2020a) ou ajout d'étiquettes de genre (Saunders *et al.*, 2020b)) pour diminuer les biais de genre tout en gardant des traductions de qualités.

Aujourd'hui l'écriture inclusive est de plus en plus utilisée, bien que sa complexité soit encore critiquée. La norme volontaire NF Z71-300 publié par l'AFNOR au début d'avril 2019 (AFNOR, 2019) qui propose une nouvelle disposition des caractères afin de mettre à jour les claviers avec notamment l'ajout du point médian pourrait permettre une plus grande diffusion de l'usage de ce symbole et de manière globale soutenir la popularisation de l'écriture inclusive.

RÉFÉRENCES BIBLIOGRAPHIQUES

Abeillé, A., An, A., & Shiraïshi, A. (2018). L'accord de proximité du déterminant en français.

Discours, 22. <https://journals.openedition.org/discours/9542>

AFNOR. (2019, avril 2). Clavier français

- Evain, A. (2008). Histoire d'autrice, de l'époque latine à nos jours. *SÊMEION, Femmes et langues*(6).
- Fontenelle, T. (2005). Dictionnaires et outils de correction linguistique. *Revue française de linguistique appliquée*, X(2), 119-128. Cairn.info. <https://doi.org/10.3917/rfla.102.0119>
- Gabriel, U., & Gygax, P. (2016). Gender and Linguistic Sexism. In *Advances in intergroup communication* (Peter Lang Publishing Inc, p. 355).
- Gabriel, U., Gygax, P., Sarrasin, O., Garnham, A., & Oakhill, J. (2008). Au pairs are rarely male : Norms on the gender perception of role names across English, French, and German. *Behavior research methods*, 40, 206-212. <https://doi.org/10.3758/BRM.40.1.206>
- Garrat-Valcarcel, R. (2020, mai 6). Coronavirus : Les Français ont pris en moyenne 2,5 kg pendant le confinement. *20minutes*. <https://www.20minutes.fr/sante/2773847-20200506-coronavirus-francais-pris-moyenne-25-kg-pendant-confinement>
- Grefenstette, G., & Tapanainen, P. (1994). *What is a word, What is a sentence*

assistons-a-une-volonte-deliberee-de-nous-empecher-de-travailler-plus-de-300-journalistes-denoncent-les-violences-policieres_3416561.html?

fbclid=IwAR06yGPmWezdxa84gl9wVFHqy-PTiOa1ILgMucXX-
pqAQs4dYbq9hc0sY_g#xtor=CS2-765-%5Bfacebook%5D-

Justine. (2020, août 28). Comment choisir son ou sa psy ? *madmoizelle*.

<https://www.madmoizelle.com/comment-choisir-son-psy-514143>

Lebarbé, T. (2015). TAL & syntaxe objets, objectifs, ambitions et nouveaux défis. *Éla. Études de linguistique appliquée*, 180(4), 463-473. Cairn.info. <https://doi.org/10.3917/ela.180.0463>

Leppihalme, R. (1997). *Culture Bumps : An Empirical Approach to the Translation of Allusions* *Topics in Translation* 10 (Multilingual Matters).

Mathot, R. (2018, mars 9). Florence Montreynaud : « Même la grammaire a un sexe. Et il est masculin depuis trois siècles... » *Levif*. https://www.levif.be/actualite/international/florence-montreynaud-meme-la-grammaire-a-un-sexe-et-il-est-masculin-depuis-trois-siecles/article-normal-809999.html?cookie_check=1596531105

Owji, Z. (2013). Translation Strategies : A Review and Comparison of Theories. *Translation Journal*, 17(1). <http://translationjournal.net/journal/63theory.htm>

Poibeau, T. (2020, janvier 23). *Babel 2.0. Où va la traduction automatique ?* [Conférence]. Petits Déjeuners de la Science et de l'Innovation, Paris.

<https://www.industrienationale.fr/evenement/babel-20-ou-va-la-traduction-automatique>

Prates, M. O. R., Avelar, P. H. C., & Lamb, L. C. (2018). Assessing Gender Bias in Machine Translation—A Case Study with Google Translate. *Neural Computing and Applications*, 1-19.

RFI. (2019, décembre 9). En Finlande, Sanna Marin devient Première ministre à seulement 34 ans. *RFI*. <https://www.rfi.fr/fr/europe/20191209-finlande-sanna-marin-plus-jeune-premiere-ministre>

Saunders, D., & Byrne, B. (2020a). *Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

- Saunders, D., Sallis, R., & Byrne, B. (2020b). *Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It. Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 35-43.
<https://www.aclweb.org/anthology/2020.gebnlp-1.4/>
- Schmidt, S. (1995). Inclusive Language and Translations. *SWET Neswletter*, 64.
https://swet.jp/articles/article/inclusive_language_and_translations/_C32
- Sebagh, C., & Baric, C. (2019). *Manuel d'écriture inclusive*. Mots-Clés. <https://www.ecriture-inclusive.fr/>
- Spriet, T., Béchet, F., El-Bèze, M., De Loupy, C., & Khouri, L. (1996). *Traitement Automatique des mots inconnus*. 170-179.
https://www.researchgate.net/publication/229039693_Traitement_automatique_des_mots_inconnus
- Sun, T., Gaut, A., Tang, S., Huang, Y., Elshierief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). *Mitigating Gender Bias in Natural Language Processing : Literature Review. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630-1640. <https://www.aclweb.org/anthology/P19-1159/>
- Tibblin, J. (2019). *Le langage inclusif : Attitudes et évaluation de texte* [Mémoire, Université de Lund]. edsbas. <http://scd-rproxy.u-strasbg.fr/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsbas&AN=edsbas.D507896F&lang=fr&site=eds-live&scope=site>
- Tribune collective. (2020, juin 2). Nous demandons l'ouverture du RSA pour les moins de 25 ans. *l'Humanité*. <https://www.humanite.fr/nous-demandons-louverture-du-rsa-pour-les-moins-de-25-ans-689831>
- Un collectif d'ancien·ne·s et actuel·le·s étudiant·e·s des grandes écoles de commerce françaises. (2020, janvier 16). «Nous aussi» : Lettre contre le sexisme, l'homophobie et le racisme dans les grandes écoles de commerce. *Libération*.
https://www.liberation.fr/debats/2020/01/16/nous-aussi-lettre-contre-le-sexisme-l-homophobie-et-le-racisme-dans-les-grandes-ecoles-de-commerce_1773224?

fbclid=IwAR1cK8LbVUbfDITVJQXKoz9xxzNgQjbmBhljMNwyTYdFXu-
sNSCoWJRuMaw

Ungerleider, N. (2013, mai 28). Google Translate's Gender Problem (And Bing Translate's, And Systran's...). *FastCompany*. <https://www.fastcompany.com/3010223/google-translates-gender-problem-and-bing-translates-and-systrans>

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). *Getting Gender Right in Neural Machine Translation*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. <https://www.aclweb.org/anthology/D18-1334>

Viennot, E. (2018). *Le langage inclusif : Pourquoi, comment ?* (iXe éd.).

Viennot, E., Candea, M., Chevalier, Y., Duverger, S., & Houdebine, A.-M. (2016). *L'Académie contre la langue française : Le dossier « féminisation »* (iXe éd.).

Winckler, M. (2019, novembre 19). La levée du secret professionnel est une mesure inappropriée, paternaliste et dangereuse pour les femmes. *Huffingtonpost*.

https://www.huffingtonpost.fr/entry/la-levee-du-secret-professionnel-est-une-mesure-inappropriee-paternaliste-et-dangereuse-pour-les-femmes_fr_5dd3afe9e4b082dae8137cfc?ncid=other_huffpostre_pqylmel2bk8&utm_campaign=related_articles

1. Abréviations utilisées

Corpus

Texte 1 : (Justine, 2020)

Texte 2 : (Boukobza, 2020)

Texte 3 : (Journalistes ayant signé la tribune, 2019)

Texte 4 : (RFI, 2019)

Texte 5 : (Un collectif d'ancien·ne·s et actuel·le·s étudiant·e·s des grandes écoles de commerce françaises, 2020)

Texte 6 : (Winckler, 2019)

Texte 7 : (Garrat-Valcarcel, 2020)

Texte 8 : (Tribune collective, 2020)

Stratégies discursives

DFA : Doubles formes abrégées

DFI : Doubles formes intégrales

RP : Règle de proximité

CL: Créations langagières

EFE : Épicènes et formulations englobantes

Stratégies de traduction

O : Omission

M : Maintien

R: Remplacement

A : Ajout