Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

La traduction littéraire automatique : Adapter la machine à la traduction humaine individualisée

Damien Hansen [1,2], **Emmanuelle Esperança-Rodier** [1], **Hervé Blanchon** [1], **Valérie Bada** [2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
[2] Université de Liège, CIRTI**, 4020 Liège, Belgique

Auteur de correspondance : Damien Hansen , prénom.nom@uliege.be

Résumé

La traduction automatique neuronale et son adaptation à des domaines spécifiques par le biais de corpus spécialisés ont permis à cette technologie d'intégrer bien plus largement qu'auparavant le métier et la formation des traducteur·trice·s. Si le paradigme neuronal (et le *deep learning* de manière générale) a ainsi pu investir des domaines parfois insoupçonnés, y compris certains où la créativité est de mise, celui-ci est moins marqué par un gain phénoménal de performance que par une utilisation massive auprès du public et les débats qu'il génère, nombre d'entre eux invoquant couramment le cas littéraire pour (in)valider telle ou telle observation. Pour apprécier la pertinence de cette technologie, et ce faisant surmonter les discours souvent passionnés des opposant·e·s et partisan·e·s de la traduction automatique, il est toutefois nécessaire de mettre l'outil à l'épreuve, afin de fournir un exemple concret de ce que pourrait produire un système entraîné spécifiquement pour la traduction d'œuvres littéraires. Inscrit dans un projet de recherche plus vaste visant à évaluer l'aide que peuvent fournir les outils informatiques aux traducteurs et traductrices littéraires, cet article propose par conséquent une expérience de traduction automatique de la prose qui n'a plus été tentée pour le français depuis les systèmes probabilistes et qui rejoint un nombre croissant d'études sur le sujet pour d'autres paires de langues. Nous verrons que si les résultats sont encourageants, ceux-ci laissent présager une tout autre manière d'envisager la traduction automatique, plus proche de la traduction humaine assistée par ordinateur que de la post-édition pure, et que l'exemple des œuvres de littérature soulève en outre des réflexions utiles pour la traduction dans son ensemble.

Mots-clefs

traduction littéraire automatique ; traduction littéraire assistée par ordinateur ; traduction automatique neuronale ; adaptation au domaine ; *fantasy* ; littérature de jeunesse ; figures du traducteur

I INTRODUCTION

1.1 L'avènement du *deep learning*

Il est difficile, depuis quelques années, de passer à côté des gros titres renvoyant aux progrès « considérables » de la traduction automatique (TA) ou plus largement, puisqu'on la désigne de plus en plus volontiers par cette synecdoque, de l'intelligence artificielle (IA) et du *deep learning*. Face à cette présence accrue de la TA, élément récurrent également dans nombre de colloques en traductologie, la créativité et l'image traditionnelle du traducteur comme passeur de culture [Lee-Jahnke, 2005] se réaffirment plus que jamais aujourd'hui en tant que plus-value de la « biotraduction » [Froeliger, 2013, Loock, 2019].

∗ Institute of Engineering Univ. Grenoble Alpes
∗∗ Centre Interdisciplinaire de Recherche en Traduction et en Interprétation

Objectivement, d'ailleurs, ce qui marque surtout la traduction automatique neuronale (TAN) vis-à-vis des paradigmes qui l'ont précédée, c'est peut-être moins un saut exceptionnel de performance que le fait d'avoir atteint un niveau suffisant pour pouvoir être utilisée couramment par le plus grand nombre, de même que sa formidable présence dans les médias. D'autre part, les discours qu'elle génère se distinguent quant à eux par la polarisation des débats et par les exagérations notables de part et d'autre de ces discussions [Loock, 2020, Cambreleng, 2020]. Les titres accrocheurs relevés dans la presse ne sont évidemment pas chose nouvelle et se retrouvent parfois presque mot pour mot dans des coupures de journaux dont Hutchins [2004] notait déjà le caractère spéculatif et optimiste à la suite de l'expérience de Georgetown en 1954. Il s'agit là, en effet, d'une tradition au moins aussi vieille que le Turc mécanique du XVIII^e^ siècle. Toutefois, le regain d'intérêt récent pour la TA est notable en ce sens qu'il s'est étendu au-delà du seul champ des traducteur·trice·s et informaticien·ne·s pour investir la sphère publique. Ce battage médiatique autour de la TA, tout comme d'autres applications issues du *deep learning*, s'explique ainsi en partie par le rôle qu'ont joué les réseaux sociaux dans la diffusion de cet engouement pour l'IA et la TA, mais aussi par le partage de démos comme *Philosopher AI* ou *Dungeon AI*, qui visaient à montrer que la machine pouvait à présent se tailler une place dans des domaines insoupçonnés, y compris créatifs. Plus important encore, l'ampleur du phénomène tient avant tout au fait que ces applications, et la TA en particulier, sont désormais des produits commerciaux portés par des entreprises qui distribuent ces outils et des personnes qui voient dans l'IA des intérêts personnels ou financiers [Rao, 2020].

Or, on peut constater à ce titre que les évocations de la TA sont le plus souvent loin d'être neutres et qu'elles sont empreintes de phénomènes discursifs forts. Parmi les éléments relevés, Rossi [2019] fait remarquer que, outre la personnification de la machine présentée non plus comme un outil mais comme « un traducteur » élevé au même rang que l'humain, celle-ci lui est toujours opposée. L'auteure note par ailleurs que, sous couvert de la mention d'intelligence artificielle, cette technologie nous est présentée comme une machine douée d'intuition, capable de généraliser et de développer de véritables connaissances, parfois même sans aucune forme d'apprentissage et bien entendu sans rappeler que les ressources qui lui permettent d'atteindre ces performances sont produites par des traducteur·trice·s professionnel·le·s. Tout ceci contribue au bout du compte à véhiculer une image de la TA qui la rapproche plus des machines à langage fictionnelles de Landragin [2020] et donc d'une IA « forte », autrement dit un système capable d'apprendre tout type de tâche de façon autonome, plutôt que des IA dites « faibles » qui existent actuellement, à savoir des machines entraînées par l'humain pour effectuer une tâche spécifique dans un environnement contrôlé. Il est dès lors important de promouvoir, à l'inverse, des approches plus raisonnées et objectives par un usage critique de la traduction automatique, à l'image du concept anglophone de *MT literacy* [Bowker et Ciro, 2019, Rossi, 2019].

Face aux discours les plus optimistes concernant la TA, les textes littéraires sont régulièrement évoqués en contre-exemple, voire, dans les cas extrêmes, pour rejeter en bloc cette technologie. Cette opposition de la littérature comme figure d'exception face aux outils informatiques n'a, elle non plus, rien de nouveau [Hansen, à paraître], mais elle semble plus marquée dans ce cas, précisément en raison de la proportion et du caractère tranché des débats sur la traduction automatique dans son ensemble. Les systèmes existants, habituellement entraînés sur des corpus institutionnels, des données issues du Web ou des textes de presse, n'ont pourtant pas été prévus pour traduire des textes littéraires. Plus encore, l'idée prévalente de l'incompatibilité entre la littérature et la machine a dissuadé les ingénieur·e·s de simplement tenter l'expérience, l'objection préjudicielle se renouvelant ainsi sans que l'on puisse compter sur des essais de traduction littéraire automatique (TLA).

Si l'on veut évaluer les performances de la TA en littérature, il est évidemment nécessaire de poser un regard objectif sur la question et de mettre les outils à l'épreuve, comme le propose depuis peu l'Observatoire de la traduction automatique [1]. Il est aussi essentiel de livrer une vision plus réaliste de la TA et de l'IA de manière générale, soit des algorithmes non dotés de conscience qui ne peuvent faire montre d'une intelligence ou de compétences comparables à celles de l'humain et dont le fonctionnement consiste essentiellement à apprendre et à répéter des schémas tirés d'un corpus d'apprentissage [Moorkens et al., 2018]. Cette conception pragmatique, ou plutôt mathématique, de la traduction automatique n'invalide pas pour autant les progrès récemment accomplis dans le domaine [Zydroń, 2019]. Au contraire, cela signifie, d'une part, que ces systèmes ont leurs limites, notamment liées aux ambiguïtés linguistiques, à leur dépendance à l'égard des données d'entraînements, aux coûts de traitement, etc. ; de l'autre, qu'ils reposent sur des postulats fondamentaux qui ne rendent pas tout à fait inenvisageable leur application dans des domaines plus créatifs. Le premier et principal de ces fondements est la nécessité d'entraîner les systèmes neuronaux sur des données du domaine en question.

1.2 Des outils (in)utiles en littérature ?

Dans un contexte d'intérêt croissant pour l'aide que peuvent apporter les nouvelles technologies (au sens large) dans le domaine littéraire [Hansen, à paraître], plusieurs auteur·e·s ont ainsi bravé l'anathème et se sont lancé·e·s dans des expériences visant à explorer la possibilité d'adapter des systèmes de traduction automatique au domaine littéraire. De même, et comme nous l'avons fait pour la traduction littéraire assistée par ordinateur (TLAO), dont l'histoire et la trajectoire ne sont pas entièrement incompatibles avec celle de la TLA [2], notre objectif est donc d'examiner la pertinence de la TA envisagée en tant qu'outil des traducteur·trice·s littéraires, mais aussi les avantages et les inconvénients que pourrait apporter l'arrivée cette technologie en littérature. Si l'on s'en tient aux discours traditionnels sur le sujet, le champ représente d'ailleurs un défi particulier qui présenterait cet avantage de mettre clairement en lumière à la fois les avancées et les limites de la TA, mais aussi la plus-value de la traduction humaine vis-à-vis de la machine, et ce, peu importe le domaine considéré.

Pour ce faire, nous avons mis au point un système de traduction automatique spécialisé non pas uniquement sur des données littéraires, mais plus spécifiquement sur la production d'une auteure et d'une traductrice dans le genre de l'*heroic fantasy*. Notre approche n'est dès lors pas celle d'un système capable de traduire une œuvre dans son intégralité en remplaçant l'humain, mais plutôt d'un outil qui l'assisterait durant le processus de traduction et qui s'adapterait à sa voix et à son style. Elle se rapproche en ce sens du paradigme avancé dans Mirkin et al. [2015] ou Mirkin et Meunier [2015], avant même l'apparition de fonctionnalités telles que « AdaptiveMT » ou « ModernMT ». La méthodologie employée pour élaborer ce système est décrite à la suite de cette introduction, après un bref état de l'art, de même que les premiers résultats obtenus lors de cette étude. Nous terminerons ensuite avec une discussion concernant ces sorties de traduction automatique et ce qu'elles laissent présager, ainsi qu'avec une courte réflexion sur la complexité des textes littéraires considérée du point de vue de la machine et, enfin, nos perspectives de recherche futures pour ce projet.

1. https://www.atlas-citl.org/lobservatoire-de-la-traduction-automatique/.

2. Cela vaut par exemple pour les arguments généralement opposés à ces technologies, que ce soit leur impossibilité de restituer la poétique d'une œuvre ou encore le contraste absolu entre la mécanicité de l'informatique et la démarche créative. À l'inverse, cela concerne aussi l'aide que pourraient offrir ces différents programmes s'ils sont utilisés avec des corpus adéquats ou, mieux encore, s'ils étaient véritablement pensés ou remaniés pour un usage littéraire.

Finalement, nous verrons que même s'il reste du chemin à parcourir, l'idée d'un moteur de traduction automatique personnalisé paraît non seulement envisageable, mais aussi profitable. Ces premiers résultats indiquent également que si la littérature accentue les faiblesses de la TA, celles-ci ne sont pas spécifiques au champ littéraire. Sans surprise, il apparaît en effet que la technologie présente des lacunes qui nous sont bien connues aujourd'hui et que l'intervention humaine reste nécessaire, peu importe le domaine. Contrairement aux machines de la science-fiction, les systèmes de TA ne peuvent de toute évidence comprendre les textes soumis ou les traductions restituées, ni même créer quelque chose de tout à fait original, bien qu'ils soient bel et bien capables de reproduire ce qui a préalablement été observé et donnent parfois lieu à des productions étonnantes ou même créatives. Le dialogue et l'interaction humain-machine, en revanche, font intervenir des compétences qui, quoique tout à fait opposées, n'en sont pas moins complémentaires et peuvent permettre de renforcer la prise de décision humaine et la démarche créative.

II ÉTAT DE L'ART

2.1 La traduction automatique neuronale : une question d'adaptation

Si l'idée d'adapter la TA à la littérature s'est posée pour les systèmes de traduction automatique probabilistes, notamment dans les travaux de Besacier [2014] ainsi que de Toral et Way [2015a], c'est principalement avec l'apparition des systèmes neuronaux que la question se fera plus insistante. Parmi les points forts avancés de la TAN [Sutskever et al., 2014, Cho et al., 2014], et plus particulièrement des modèles neuronaux avec mécanisme d'attention [Bahdanau et al., 2014, Luong et al., 2015], il avait effectivement été remarqué que ces systèmes produisaient des traductions moins littérales et plus naturelles, qu'ils étaient plus efficaces avec des textes lexicalement riches et qu'ils commettaient moins d'erreurs lexicales et morphologiques [Bentivogli et al., 2016]. Chacun de ces points pouvait donc laisser augurer de meilleurs résultats pour la traduction d'ouvrages littéraires, mais une nouvelle architecture de réseau de neurones centrée entièrement sur ces mécanismes d'attention est rapidement venue rehausser d'un cran supplémentaire la qualité des traductions. De fait, ces modèles de traduction automatique *Transformer* produisent de meilleurs résultats encore, car il a été remarqué entre autres choses qu'ils pouvaient capturer des dépendances linguistiques à plus grande échelle [Vaswani et al., 2017]. Ce sont ces mêmes modèles qui ont véritablement lancé les discussions sur la TA en traductologie et qui ont aussi attiré une très vive attention du public, puisqu'ils sont à la base des modèles de langues GPT-2 et GPT-3, par exemple, qui ont été abondamment diffusés dans la presse et sur les réseaux sociaux.

Néanmoins, une particularité de l'approche neuronale est qu'elle est relativement gourmande en données et qu'elle nécessite des corpus d'entraînement issus du domaine dans lequel elle sera déployée pour atteindre la meilleure performance [Chu et Wang, 2018], faisant de l'adaptation au domaine l'un des principaux défis posés par la TA encore aujourd'hui [Müller et al., 2020]. De nombreuses méthodes ont ainsi été proposées pour adapter la TAN à des langues ou à des domaines spécifiques pour lesquels peu de ressources étaient disponibles, permettant à la TA de s'étendre vers de nouveaux secteurs. Évidemment, le cas idéal est celui où le système peut être entraîné uniquement sur des données *ad hoc*, mais ce scénario n'est pas toujours réalisable en pratique et représente à ce titre un enjeu fondamental de la traduction littéraire automatique. Dans tous les cas, l'élaboration et la mise à disposition de moteurs de traduction adaptés à des domaines spécifiques est devenue pratique courante et est d'ailleurs l'un des principaux arguments de vente des prestataires de service aujourd'hui.

2.2 Vers une traduction littéraire automatique ?

Ce constat lié au besoin de spécialisation rejoint les conclusions tirées par Besacier [2014], Toral et Way [2015a], dont les deux cas d'étude pour les paires de langues anglais-français et espagnol-catalan mettaient déjà en lumière la nécessité d'adapter leur système de traduction automatique sur des données littéraires. Couplée avec l'arrivée de la TAN, l'idée de mettre au point des systèmes de TA spécifiques à la littérature semble s'être répandue depuis lors. Depuis 2018, on constate de ce fait quelques mises à l'essai dont nous donnons un aperçu plus détaillé dans Hansen [2021] et dont on peut en outre trouver des équivalents pour la traduction et la génération automatique de poésie [Ghazvininejad et al., 2018, Van de Cruys, 2019]. Elles s'ajoutent à d'autres études qui cherchent à évaluer les performances de moteurs publiquement accessibles comme *Google Traduction* ou *DeepL*, permettant ainsi de voir la qualité des textes produits par les outils qui sont utilisés au quotidien par le plus grand nombre. Il faut toutefois bien garder à l'esprit que ces systèmes n'ont pas du tout été prévus dans ce sens et qu'il est injuste de rejeter totalement la possibilité de la TLA en se basant uniquement sur ces outils.

Parmi les travaux cherchant à mettre au point un moteur de traduction littéraire automatique, Toral et Way [2018] montrent que l'on peut obtenir des résultats encourageants avec des systèmes de TAN entraînés sur des données uniquement littéraires. Les chercheurs utilisent pour ce faire un volumineux corpus de livres électroniques, composé de 133 romans en anglais alignés avec leur traduction vers le catalan (1 million de phrases) et d'un sous-corpus synthétique [3] obtenu à partir de 1 000 romans en catalan (5 millions de phrases). La constitution d'un tel jeu de données n'est cependant pas anodine et très peu d'autres corpus existent pour la littérature. Pour cette raison, Matusov [2019] et Kuzman et al. [2019] suggèrent de concevoir des systèmes de TA génériques et de les affiner ensuite sur des textes littéraires. Dans le premier cas, l'auteur élabore un moteur anglais-russe auquel sont fournies des données hors domaine combinées à un corpus du domaine (270 000 phrases) et à un autre synthétique (2,3 millions de phrases), tandis qu'un deuxième moteur de traduction allemand-anglais est conçu de façon similaire (50 000 phrases et 10 millions de phrases). Les résultats montrent que ce processus d'adaptation produit une légère amélioration pour le russe et une tendance inverse dans l'autre scénario, bien que les évaluations humaines témoignent d'un gain de qualité pour les deux systèmes selon le chercheur. Dans le second cas, c'est un moteur de TA anglais-slovène qui est entraîné sur des données hors domaine et affiné par la suite sur 9 romans, montrant lui aussi un léger gain de performance par rapport à celui construit seulement sur les données génériques.

III MÉTHODOLOGIE

3.1 Un système de traduction automatique personnalisé

Notre étude se rapproche des deux dernières décrites ci-avant, dans la mesure où nous disposons de peu de données littéraires et où nous tentons d'affiner un système générique pour ce domaine. Plus encore, l'objectif sous-jacent de cette expérience est d'explorer également la possibilité d'adapter ce moteur de TA au travail du traducteur ou de la traductrice. Elle fait ainsi écho aux remarques de Mirkin et al. [2015], qui proposaient de voir quels pourraient être les avantages d'un système personnalisé capable de conserver des traits d'auteur autrement gommés par la TA, envisageant dès lors une tâche d'adaptation centrée non pas sur un domaine, mais sur l'humain.

3. Les corpus synthétiques sont des jeux de données monolingues récoltés dans la langue cible et traduits automatiquement pour produire une langue source artificielle. Cette méthode avancée par Sennrich et al. [2016] permet de constituer rapidement des corpus parallèles conséquents qui améliorent la qualité des sorties de traduction.

Cette mise à l'essai répond en outre à la principale observation tirée par Kuzman et al. [2019], qui révèle qu'un système affiné simplement sur une traduction de la même auteure et de la même traductrice donne de meilleurs résultats que ceux produits par un autre moteur entraîné sur un corpus plus vaste et plus varié d'ouvrages littéraires. Enfin, cet article présente une nouvelle tentative de traduction littéraire automatique depuis l'anglais vers le français qu'il ne nous a plus été donné de voir depuis les travaux de Besacier [2014] sur la TA probabiliste.

Pour observer concrètement les performances de la TA dans ce scénario, nous sommes donc repartis d'un corpus précédemment utilisé lors d'une expérience de TLAO. Cette mémoire de traduction, qui nous permet à présent d'établir un contraste avec la TLA, avait été constituée semi-manuellement et compilée à partir de la série d'*heroic fantasy* intitulée *Septimus Heap*. Cette saga littéraire à succès, écrite par Angie Sage (HarperCollins, 2005–2013) et traduite par Nathalie Serval (Albin Michel, 2005–2013), est composée de 7 tomes, dont le dernier n'est finalement jamais paru en français. Elle nous semblait dès lors constituer un corpus idéal dont la forte cohérence interne pourrait équilibrer, à l'inverse, la taille très limitée des données. En effet, ce jeu de données que nous décrivons à la suite paraît bien peu de chose quand on sait qu'un jeu d'entraînement de 6 millions de segments est considéré comme « frugal » pour mettre au point un système de TA aujourd'hui [Blin, 2021]. La constitution d'un moteur de TA générique à affiner sur des données plus restreintes a par conséquent été une étape indispensable du processus, d'autant plus que très peu de corpus littéraires sont disponibles en ligne, mais elle nous a donné l'occasion de vérifier s'il était possible, *in fine*, d'adapter un système aux œuvres spécifiques d'une saga, d'une auteure et d'une traductrice.

3.2 Systèmes et données

Le premier obstacle posé par cette expérience a tout d'abord été, comme nous le disions, un problème de données. Comme indiqué dans le Tableau 1, notre corpus spécialisé compte au total 45 277 segments [4], qui doivent être par ailleurs séparés en un jeu d'entraînement et de validation (pour mettre au point le système), mais aussi en un jeu de test (pour évaluer ensuite ses performances sur des segments non vus au préalable). Les cinq premiers tomes de la série composent les données d'entraînement, avec 37 348 segments, tandis que les 7 225 segments utilisés comme corpus de validation sont tirés du sixième volume. Nous en avons cependant extrait les trois derniers chapitres, pour lesquels nous pouvions compter sur une traduction de référence. Ces 704 derniers segments forment notre jeu de test pour cette étude et n'interviennent pas durant l'apprentissage, étant donné que le système doit être mis à l'épreuve sur des extraits qui lui sont inconnus.

Le seul autre corpus littéraire publiquement disponible pour la paire anglais-français est, à notre connaissance, le corpus Books, du projet OPUS [5] [Tiedemann, 2012]. Néanmoins, ce jeu de données est lui aussi relativement restreint et comporte environ un tiers des fragments pour lesquels les langues sont inversées, de même que de nombreux problèmes d'alignement qui s'expliquent par la constitution automatique de ce travail déjà conséquent. Pour ces raisons, il était nécessaire d'élaborer en premier lieu un système générique que nous pourrions ensuite affiner sur ces quelques 45 000 segments, c'est pourquoi nous nous sommes tournés vers d'autres jeux de données habituellement utilisés dans le champ de la traduction automatique pour la paire anglais-français.

4. Nous parlons ici de segments et non de phrases, car certaines lignes ont été agglutinées pour accommoder d'éventuelles différences de structure entre les deux versions. Chaque ligne devant contenir exactement les mêmes informations, elles ont ainsi été combinées jusqu'à coïncider et peuvent parfois atteindre la taille d'un paragraphe.

5. https://opus.nlpl.eu/.

En plus du corpus Books, pour lequel nous avons corrigé le problème d'inversion, nous avons donc utilisé d'autres corpus librement accessibles du dépôt OPUS, à savoir Europarl (v8), GlobalVoices (2018Q4), News-Commentary (v16) et TED2020 (v1). Nous y avons ajouté un corpus personnel de traduction de jeux vidéo décrit dans Hansen et Houlmont [2022], en raison non seulement de la propreté et de la taille du matériau, provenant majoritairement de la *fantasy*, mais aussi pour équilibrer un ensemble de données autrement issues en grande partie des institutions ou de la presse. Le matériel compilé s'élève au total à un peu moins 4 millions de segments, dont nous donnons les spécifications dans le Tableau 2.

	Segments	Tokens EN	Tokens FR
Entraînement	37 348	550 536	541 779
Validation	7 225	109 859	106 621
Test	704	10 181	10 073
Total	45 277	670 576	658 473

TABLEAU 1 – Corpus spécialisé.

	Segments	Tokens EN	Tokens FR
Europarl	2 007 723	49 867 465	54 553 979
JV	956 658	14 006 849	15 222 816
TED	410 443	7 041 745	7 464 033
GlobalVoices	195 387	3 503 600	3 980 602
News	183 251	4 055 180	4 952 704
Books	127 021	2 737 133	2 770 418
Total	3 880 483	81 211 972	88 944 552

TABLEAU 2 – Corpus génériques.

Ce corpus générique contient de ce fait des genres hétérogènes, y compris des textes littéraires, car nous voulions mettre plus particulièrement en évidence les résultats liés à l'ajout de notre corpus spécialisé. Nous avons élaboré également un système auquel nous avons uniquement présenté notre corpus personnel, à titre de comparaison, pour illustrer l'importance de la taille des données dans l'entraînement d'un moteur de traduction automatique neuronal. Pour chacun de ces trois scénarios (littéraire uniquement, générique et affiné), nous avons comparé les résultats produits par un modèle LSTM [Bahdanau et al., 2014] et par un modèle *Transformer* [Vaswani et al., 2017]. Dans tous les cas, les jeux de données ont été pré-traités avec le tokeniseur Moses [Koehn et al., 2007] et segmentés avec le modèle unigram de sentencepiece [Kudo, 2018] pour aboutir en définitive à une taille de vocabulaire de 16 000 tokens, mais nous n'avons toutefois pas normalisé la casse car elle exerce un rôle déterminant dans notre série littéraire. Les six systèmes maison ont ensuite été mis au point avec la v2 du module OpenNMT [Klein et al., 2017].

Les trois modèles Bi-LSTM ont été entraînés sur un GPU avec les paramètres suivants : 3 couches avec une dimension cachée de 1024, des plongements (*embeddings*) de dimension 512, le mécanisme d'attention *general* et un apprentissage par lot (*batch size*) de 24 phrases. Les modèles *Transformer* ont tous les trois été configurés avec les paramètres par défaut de l'architecture de base décrite dans Vaswani et al. [2017], soit 6 couches avec une dimension cachée de 2048, des plongements de dimension 512, 8 têtes d'attention, un apprentissage par lot de 4096 tokens et une régularisation par abandon (*dropout*) fixée à 0,1. Les systèmes entraînés seulement sur le corpus spécialisé (« Septimus ») ont tourné pendant un total de 200 000 étapes, tout comme les systèmes confrontés uniquement aux corpus génériques (« Générique »). Les deux derniers systèmes (« Affiné ») ont été relancés à partir des modèles génériques sur 100 000 étapes supplémentaires, en conservant néanmoins les données génériques ajustées avec un poids moindre pour éviter un oubli catastrophique (*catastrophic forgetting*) au cours de l'affinage. Durant cette même phase, nous avons par ailleurs ajouté le corpus de validation aux données d'entraînement et poursuivi celui-ci sans évaluation pour que le système soit confronté au plus grand nombre d'exemples. La traduction a été réalisée quant à elle avec les paramètres par défaut d'OpenNMT, qui utilise une recherche en faisceau (*beam search*) de taille 5.

IV RÉSULTATS

4.1 Évaluations automatiques

Nous indiquons le score BLEU [Papineni et al., 2002] obtenu pour chacun des modèles dans le Tableau 3. Pour rappel, BLEU est un algorithme d'évaluation automatique qui compare la ressemblance entre un texte produit par la TA et une (ou plusieurs) traduction(s) de référence, produisant un score plus élevé lorsque le texte généré automatiquement se rapproche de la production humaine. Si l'on s'accorde généralement sur le fait que cette métrique est peu représentative de la qualité en tant que telle, elle permet cependant de se comparer avec l'état de l'art, d'apprécier l'évolution d'un système sur un jeu de test bien précis et de voir de cette manière si les modifications apportées à celui-ci conduisent ou non à une amélioration.

	Septimus	**Générique**	**Affiné**
LSTM	09.34	09.01	14.43
Transformer	09.10	09.93	18.11

TABLEAU 3 – Comparaison des modèles LSTM et *Transformer* avec différents corpus d'entraînement.

Sans trop de surprise, on peut voir ici que le modèle le plus performant est le *Transformer* et qu'il conduit à un score plus élevé lorsqu'il est entraîné à la fois sur beaucoup de données et sur des données spécialisées (modèle « Affiné »). L'architecture LSTM donne en revanche un meilleur résultat pour les systèmes entraînés uniquement sur ce corpus spécialisé (« Septimus »), confirmant l'observation selon laquelle les *Transformers* nécessitent beaucoup d'exemples pour l'apprentissage et restent moins performants sur certaines tâches lorsque peu de données sont disponibles [6], bien que le score reste globalement faible pour les deux paradigmes. Cela vaut en outre pour les deux systèmes génériques, qui peinent à dépasser un résultat supérieur à 10 dans toutes nos configurations. Seul l'ajout du corpus spécialisé sur un système déjà suffisamment robuste nous a permis de dépasser ce seuil, mais l'amélioration est dans ce cas particulièrement significative (+ 82 %). Pour les deux derniers systèmes *Transformer*, nous fournissons deux métriques supplémentaires, de même qu'une comparaison avec des traductions produites par *Google Traduction* et *DeepL* [7]. Tous ces scores ont été calculés avec sacreBLEU [Post, 2018], dont nous fournissons la signature [8], et sont compilés dans le Tableau 4.

	BLEU ⇑	**chrF2++ ⇑**	**TER ⇓**
Google Trad.	10.79	35.20	91.08
DeepL	10.04	34.88	92.81
Générique	09.93	33.14	92.24
Affiné	18.11	40.32	76.04

TABLEAU 4 – Comparaison des systèmes maison et des outils librement accessibles.

6. Nous prévoyons néanmoins de mener des essais d'optimisation pour ce scénario particulier par la suite, à la manière de Sennrich et Zhang [2019], Araabi et Monz [2020].
7. Traductions effectuées le 25/11/2020.
8. BLEU nrefs :1 | case :mixed | eff :no | tok :13a | smooth :exp | version :2.0.0
   chrF2++ nrefs :1 | case :mixed | eff :yes | nc :6 | nw :2 | space :no | version :2.0.0
   TER nrefs :1 | case :lc | tok :tercom | norm :no | punct :yes | asian :no | version :2.0.0

Pour chacune de ces métriques, une flèche indique si les performances s'améliorent à mesure que le score monte ou qu'il baisse. Ainsi, les trois évaluations confirment l'amélioration dont nous faisions part vis-à-vis du modèle générique, de même que par rapport aux systèmes de TA en ligne. Nous ne cherchons pourtant pas uniquement ici à évaluer la qualité de la traduction, mais aussi à quel point les propositions de la machine se rapprochent de celle de la traductrice française et de son style (c'est nécessaire si l'on veut que la TA soit utile aux professionnel·le·s). L'exigence est donc doublement élevée et pourrait expliquer les scores faibles des trois systèmes génériques. Or, ces trois métriques sont d'autant plus intéressantes qu'elles évaluent justement à quel point les textes produits s'écartent de la traduction de référence, mais l'on peut voir ici que l'adaptation d'un système permet tout de même d'obtenir des résultats bien plus élevés qu'à l'origine (+ 8 BLEU environ dans ce sens, avec pourtant assez peu de données et un système qui est bien loin des modèles robustes déployés en pratique par les entreprises).

Pour avoir une idée un peu plus précise de ce que représentent ces scores BLEU, nous pouvons mettre les choses en perspective de plusieurs manières. Tout d'abord, nous pouvons comparer ces résultats à ceux rapportés pour les autres adaptations littéraires, dont nous fournissons un comparatif dans le Tableau 5. Il est toutefois impératif de rappeler que BLEU est un score automatique, calculé phrase par phrase par rapport à une et une seule traduction de référence, qu'il n'est dès lors pas représentatif de la « qualité réelle » d'une sortie de TA et qu'il n'est pas non plus directement comparable d'une expérience à l'autre [9]. Un exemple lié à l'ambiguïté de BLEU dans note étude pourrait être le contraste entre les modèles « Septimus » et « Générique », qui affichent un score semblable bien que le premier produise des mots de vocabulaire correctement traduits dans des phrases tout à fait incompréhensibles, là où le second génère des phrases correctes mais totalement éloignées du champ lexical du roman [10].

	Générique	**Affiné**	**Google Trad.**
EN-RU (LSTM) [Matusov, 2019]	14.20	15.20	13.90
DE-EN (Transf.) [Matusov, 2019]	18.50	16.20	20.20
EN-SL (LSTM) [Kuzman et al., 2019]	17.50	20.75	21.97
EN-FR (Transf.) [cette étude]	09.93	18.11	10.79

TABLEAU 5 – Comparaison avec d'autres essais d'adaptation à la littérature.

Cela étant dit, nous pouvons quand même constater deux choses. La première est que l'adaptation sur des données littéraires peut parfois sembler négligeable, voire contre-productive si l'on s'en tient strictement aux métriques, mais le cas développé par Kuzman et al. [2019] offre une amélioration plus appréciable, que les chercheur·euse·s obtiennent en ajoutant aux données d'entraînement du modèle un texte écrit et traduit par les mêmes personnes que le roman évalué.

9. Un même système utilisé sur plusieurs textes peut éventuellement servir à révéler des particularités de ces textes ou à éprouver sa robustesse, surtout si l'on peut les comparer avec les données utilisées pour entraîner le système. À l'inverse, plusieurs systèmes entraînés avec différents jeux de données ou différents paramètres et confrontés à un même test montrent quelle méthode produit les meilleurs résultats (pour ce corpus de test). Cependant, comparer un score BLEU obtenu par deux systèmes fondamentalement différents sur un jeu de test distinct (et plus encore sur deux paires de langues éloignées) ne révélerait aucune sorte d'information.

10. Cette observation est d'ailleurs conforme avec l'idée que les systèmes de TA neuronaux nécessitent un grand nombre de données pour intégrer des éléments de syntaxe et un vocabulaire suffisamment riche, mais aussi des données spécialisées qui reflètent la terminologie et le style [Moslem, 2020].

La deuxième, c'est que les résultats présentés dans ces deux études restent proches ou même inférieurs à ceux donnés par *Google Traduction*, souvent utilisé comme référence. De plus, ces scores sont globalement bien plus élevés que celui donné par ce même système sur notre texte, s'élevant péniblement autour de 10 BLEU. Dans notre cas, l'entraînement sur des données du domaine, de l'auteure et de la traductrice permet pourtant de presque doubler ce score, signalant selon nous l'intérêt d'une approche visant à mettre au point des systèmes de traduction automatique personnalisés pour les traducteur·trice·s.

Pour ces quatre exemples, nous pouvons aussi noter que les évaluations donnent des résultats assez bas si on les met en perspective avec les scores BLEU entre 18 et 39 obtenus par le modèle entièrement littéraire de Toral et Way [2018], qui mentionnaient dans une précédente publication qu'un score BLEU de 20 pourrait être un point de repère donnant aux ingénieur·e·s l'indication d'une post-édition utile ou utilisable [Toral et Way, 2015b]. Cela signifie que nous n'en sommes pas encore au stade d'un outil complètement fonctionnel, et encore moins de la traduction de textes littéraires entièrement automatisée et de haute qualité, mais que l'adaptation donne tout de même des résultats très encourageants et qu'il y a du potentiel pour un outil d'aide à la traduction personnalisé. On peut toutefois se demander pourquoi il existe une telle variation dans les scores rapportés sur différentes œuvres ou paires de langues, certaines tournant autour de 10 ou 20 BLEU, d'autres montant jusque 40 et plus. Et si la prédiction de la qualité des traductions automatiques est un champ de recherche très important aujourd'hui, notamment en raison de la commercialisation de la TA, nous tenterons modestement ici d'explorer la question de la complexité des textes (littéraires) du point de vue de la machine.

4.2 Peut-on mesurer la complexité des textes littéraires (pour une machine) ?

Face à ces diverses mises à l'essai de traduction littéraire automatique, mais aussi et surtout à la difficulté de comparer ces résultats fort variables, nous nous demandons s'il serait possible d'estimer la complexité des textes littéraires et de fournir une appréciation qui expliquerait certains résultats, plutôt bons ou plutôt mauvais selon les cas étudiés et les systèmes employés. Évidemment, une telle évaluation ne pourra en rien refléter quelque chose d'aussi subjectif que la sensibilité des traducteur·trice·s pour un ouvrage en particulier, mais pourra peut-être nous renseigner sur la difficulté de leur prise en charge, d'un point de vue mathématique et statistique, par la machine.

Une piste avancée en ce sens par Toral et Way [2015a] consiste à aligner mot à mot les phrases d'un texte original et de sa traduction pour relever ensuite la perplexité affichée par l'algorithme d'alignement. L'idée est que si l'outil prévu pour cette tâche est capable d'aligner aisément les deux versions, par exemple si chaque mot du texte source correspond à un mot du texte cible, la traduction pourrait être considérée plus simple que dans le cas où ce même outil peine à trouver des correspondances et montre dès lors une plus grande mesure de perplexité. Cette méthode permettrait alors d'estimer la prise de liberté dans les traductions et, par voie de conséquence, la probabilité que les sorties de TA, généralement plus littérales, soient de bonne ou de moins bonne qualité. Dans cette même étude, les auteurs avaient trouvé que les deux romans évalués présentaient un degré de liberté moindre que leur corpus de presse, et moins encore que le corpus Europarl. En répétant l'expérience pour la paire anglais-français, nous avons obtenu des résultats qui n'étaient pas tout à fait similaires, puisque notre corpus de littérature variée (le corpus Books) révèle un indice de perplexité plus haut que pour le corpus News (32 contre 27). Et bien qu'Europarl conserve un score plus élevé (38), quoique moins éloigné que pour la paire espagnol-catalan chez Toral et Way [2015a], notre roman *Septimus Heap* obtient un résultat autrement plus élevé (55) que ce qui nous est donné de voir dans cette même étude.

Par curiosité, nous avons entrepris de reproduire l'expérience sur les différents textes littéraires écrits originellement en anglais qui composent le corpus Books. Ces indices d'alignement ont été obtenus grâce au programme GIZA++ [Och et Ney, 2003] et sont illustrés par la Figure 1.

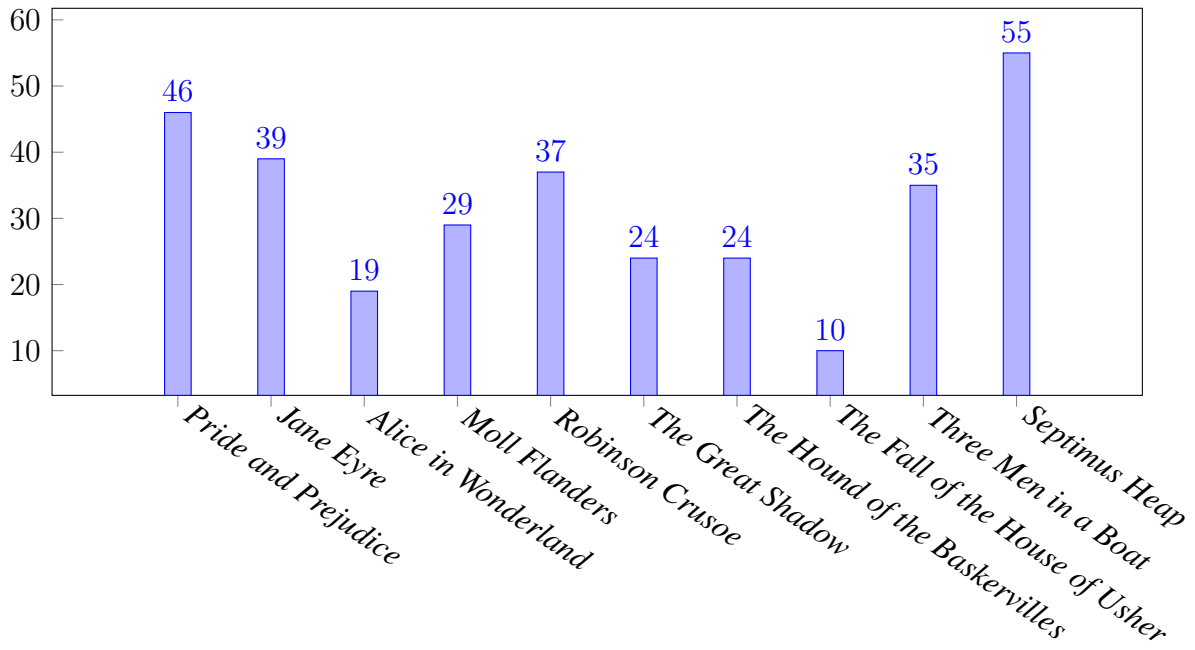

FIGURE 1 – Perplexité donnée par l'alignement des romans et de leur traduction.

Il nous faut mentionner cependant que les scores pourraient être légèrement biaisés, étant donné que ce corpus a été constitué et aligné automatiquement par la machine à l'origine. Si tel était le cas, ils devraient être plus faibles et signaler un degré de liberté inférieur, mais l'on remarque que les scores sont dans tous les cas moins élevés que pour notre ouvrage de *fantasy*, qui a été quant à lui aligné à la main et semble présenter un obstacle singulier pour la machine.

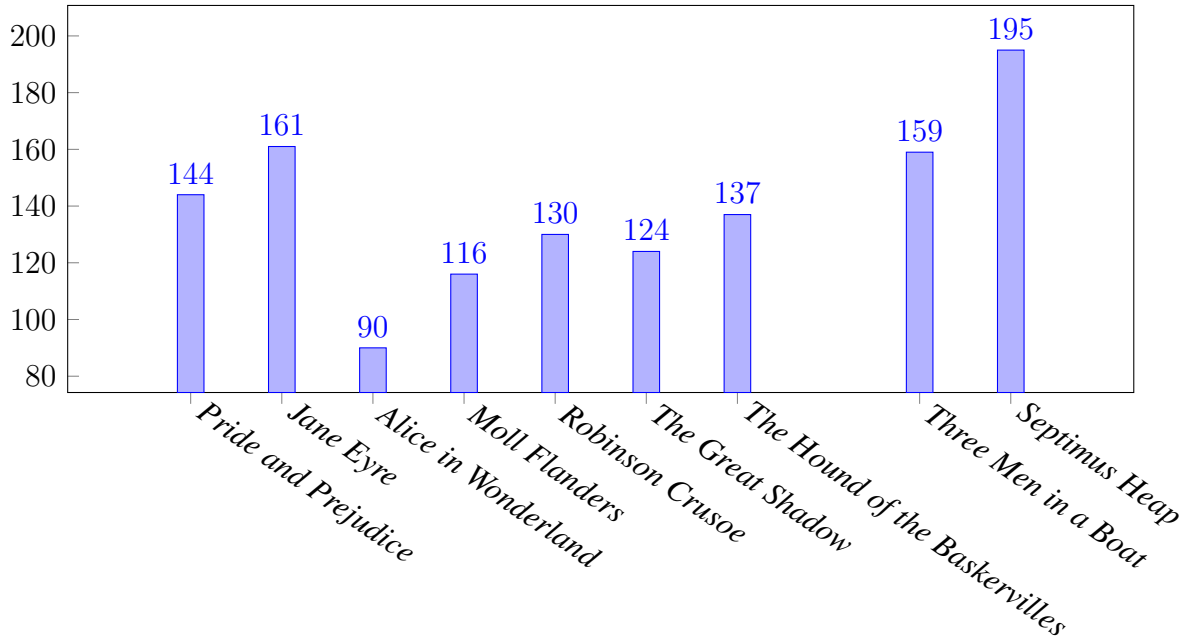

FIGURE 2 – Perplexité des modèles de langue.

Un second moyen proposé par les mêmes auteurs implique de vérifier le degré de spécialisation du domaine (*domain narrowness*) en se référant à la perplexité des modèles de langue construits sur différents corpus. En modélisant le langage de chacun des textes, il est en effet possible de calculer, une fois encore, un indice de perplexité, mais qui mesure ici l'entropie du vocabulaire utilisé et qui nous renseignerait vaguement sur le caractère prévisible ou imprévisible des textes, sur leur richesse lexicale et sur leur diversité. Plus précisément, cet indice peut être obtenu en entraînant un modèle de langue sur une section du corpus, pour laquelle un algorithme enregistre des suites d'unigrammes, de bigrammes et de trigrammes, et en lui présentant par la suite une autre section plus réduite afin de voir à quel point elle diffère de ce qui a déjà été observé. Cette façon d'envisager les choses est d'ailleurs intéressante dans la mesure où, malgré tout ce que l'on peut en dire, le *deep learning* est une approche purement statistique qui repose entièrement sur ce type de comptes et de probabilités. Nous avons donc reproduit cette opération sur notre corpus littéraire, mis à part la nouvelle d'Edgar Allan Poe pour laquelle la taille des données est insuffisante, et nous relevons dans la Figure 2 la perplexité des modèles de langue donnée dans chaque cas par l'outil SRILM [Stolcke, 2002]. Une fois encore, il semble que notre ouvrage affiche le score le plus élevé, indiquant une plus grande difficulté dans la prédiction et la modélisation de ce jeu de données.

Enfin, une dernière façon d'estimer la complexité d'une traduction automatique, qui ne nous est pas inconnue comme nous a pu le voir dans le Tableau 5, consiste simplement à passer ces documents dans un outil de TA généraliste et libre d'accès pour en tirer un score BLEU et, de cette manière, une approximation de la facilité avec laquelle ce système est capable de produire une traduction proche de la référence. En passant par un service en ligne et en calculant ce score pour plusieurs œuvres avec le même outil, on peut alors avoir une vague idée de la difficulté du texte pour la TA. Dans ce dernier cas, plus le score est faible, moins la traduction est bonne ; ou plutôt, plus elle s'éloigne de celle qui est attendue. Comme nous pouvons le voir sur la Figure 3, le roman *Septimus Heap* arrive à nouveau à la tête de ce classement [11].

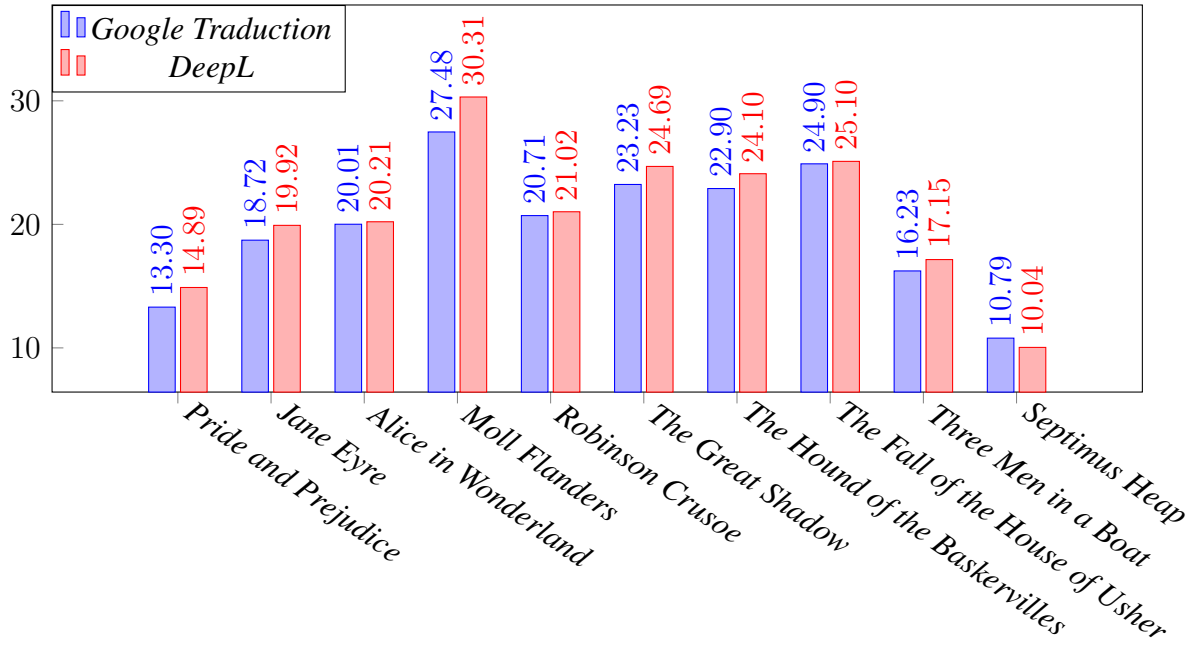

FIGURE 3 – Traduction automatique par des systèmes libres d'accès.

11. Les textes comparés ici relevant du domaine public, il n'est pas impossible que leur présence dans les données d'entraînement de *Google Traduction* et *DeepL* gonfle leur score, mais il est impossible de le vérifier.

Cette investigation demanderait bien entendu des analyses plus poussées et révèle des propriétés très différentes, mais la première chose à remarquer, dans l'ensemble, est que notre ouvrage d'*heroic fantasy* semble particulièrement complexe — il faut le rappeler — pour la machine. Ou tout du moins plus complexe que les autres textes à notre disposition et peut-être même que ceux sur lesquels la littérature s'est penchée jusqu'à présent. Nous ne prétendons pas partager ici une méthode infaillible pour mesurer la complexité des textes littéraires et il nous faut insister sur le fait que ces calculs de perplexité et de probabilité ne reflètent en rien les phénomènes qui pourraient être intuitivement complexes ou saillants pour l'humain, mais la combinaison de ces trois mesures nous semble donner une bonne indication de départ sur ce point. Notre corpus étant le plus propre, du fait qu'il a été aligné manuellement, nous aurions pourtant pu nous attendre à des résultats mitigés, mais ceux-ci se confirment pour les trois tests effectués. Une telle conclusion remet dès lors en question l'idée préconçue et relativement répandue selon laquelle la « paralittérature », la littérature de jeunesse ou la fiction, par leur prétendue simplicité stylistique, lexicale et syntaxique, seraient des cibles de choix pour la traduction automatique.

En réalité, ce constat n'a rien de tellement surprenant, étant donné que la fiction amène son lot de problèmes par-dessus ceux inhérents au domaine littéraire. Dans le cas qui nous occupe, cela peut être attribué notamment à une grande prise de liberté dans la traduction, mais aussi à un registre globalement soutenu qui peut néanmoins varier fortement d'un personnage à l'autre, à l'usage délibéré de régionalismes et d'archaïsmes, voire de discours adaptés du vieux français, ou encore à la création d'un univers fictif qui implique par conséquent l'invention de nouveaux termes, concepts et néologismes spécifiques à la saga. Ce dernier point est peut-être le plus important, car si certaines œuvres du canon littéraire sont réputées pour leur complexité, du point de vue de l'humain, elles font tout de même usage d'un vocabulaire commun, en ce sens qu'il renvoie à des termes ou à des concepts que l'on peut rencontrer ailleurs et que la machine est plus susceptible d'avoir déjà observés. À l'inverse, les textes de fictions ont une disposition mythopoétique qui se manifeste souvent par la création d'un monde, d'une histoire, d'une cosmogonie et surtout d'un vocabulaire qui leur sont propres. Ces irrealia [Loponen, 2009] dénotent un monde fictionnel et textuel qui n'a aucune existence en dehors des pages du livre et auquel les systèmes de TA n'ont probablement aucune chance d'avoir été confrontés auparavant, à moins de ne les avoir entraînés à partir d'œuvres similaires.

Or, nous voyons bien là toute la difficulté de comparer des systèmes de TA si différents, puisque mis à l'essai sur des paires de langues variables et des corpus d'entraînement variés, mais aussi appliqués à la traduction de divers·es auteur·e·s, traducteur·trice·s et genres littéraires. Cela peut également expliquer le fait que l'on puisse trouver des résultats parfois très éloignés selon l'ouvrage traduit. Contrairement à ce que l'on pourrait penser, en revanche, la littérature de jeunesse et les difficultés inhérentes à ce domaine semblent présenter une complexité toute particulière pour la traduction littéraire automatique. Notre expérience montre cependant que malgré cette situation de départ visiblement défavorable, l'adaptation à ce domaine spécifique permet d'obtenir de bien meilleurs résultats. Dans ces circonstances, et dans un contexte où l'outil continue d'évoluer au jour le jour et où d'autres chercheur·cheuse·s rapportent d'autres conclusions encourageantes, la TA se profile ainsi comme une aide intéressante, mais surtout si le système est entraîné directement par les professionnel·le·s sur leurs propres productions. Nos résultats laissent paraître selon nous une autre façon d'envisager la TA, plus proche du travail de l'humain, de son style et de ses traductions. S'il devenait possible à l'avenir d'équiper ces professionnel·le·s avec de tels outils adaptatifs, le système serait alors capable de faire des choix plus pertinents, de fournir des suggestions utiles à l'humain et peut-être, qui sait, de se tailler un jour une place dans l'arsenal des traducteur·trice·s littéraires.

V DISCUSSION

5.1 Évaluer les sorties de traduction

Pour l'heure, une autre question qui se pose actuellement, pour l'ensemble de la recherche sur la traduction automatique mais d'autant plus pour ce jeune champ de recherche, reste de voir comment évaluer au mieux les performances de la machine et, de la même manière, comment apprécier la qualité des systèmes adaptés la littérature. Nous avons eu recours dans la section précédente à plusieurs métriques, mais, si même nous aurions pu partager d'autres exemples ou d'autres types de mesures plus spécifiquement axées sur certaines caractéristiques des textes littéraires, les évaluations automatiques ne sont pas nécessairement représentatives de la qualité, comme nous l'avons évoqué, et elles n'aident pas forcément les lecteur·trice·s à se représenter ce qui est produit concrètement ou bien quels sont les points forts et points faibles de ces systèmes. Pour la suite des travaux, nous nous attacherons donc à étudier plus précisément ce volet de recherche par le biais d'un travail d'annotation d'erreur, d'enquêtes et de commentaires critiques de l'ouvrage considéré.

À ce stade, quelques observations de surface nous ont par exemple permis de constater que notre système avait tendance à traduire de façon plutôt littérale et qu'il produisait surtout des erreurs liées à de mauvais choix lexicaux ou à l'utilisation des déterminants et des propositions. Il s'agit là malheureusement de limites bien connues de la TA qui ont d'ores et déjà été observées depuis un certain temps dans d'autres domaines. Au contraire, il est surprenant de voir que le vocabulaire spécifique de la série a été scrupuleusement respecté, tout comme les conventions typographiques des précédents ouvrages, et que certains termes ont été rendus par des paraphrases astucieuses. Plus encore, le fait que certains segments comportent plusieurs phrases nous a permis d'observer que le système fusionnait certaines d'entre elles lors du passage vers le français, et que ces occurrences correspondaient toujours à une décision similaire de la part de la traductrice. Évidemment, ces quelques points sont donnés ici à titre d'exemple et devront assurément être approfondis, mesurés et illustrés plus finement dans de prochains travaux.

5.2 Évaluer les implications

L'introduction de cette technologie dans le domaine littéraire suscite en outre de nombreuses questions éthiques et pratiques dont les lecteur·trice·s intéressé·e·s trouveront un aperçu dans Hansen [2021]. Dans les conditions actuelles, une technologie telle que la TLA servirait avant tout d'aide aux apprenant·e·s ou à la circulation et à la reconnaissance des œuvres, mais elle pourrait devenir à l'avenir un véritable outil individualisé des traducteur·trice·s, si les performances des systèmes continuaient d'évoluer vers un mieux et qu'il était possible de prévoir une mise en place et une intégration aisée au sein des interfaces de travail existantes. Sans surprise, la plupart des avantages liés à l'utilisation de la TA seraient à envisager sous l'angle d'un système qui assisterait les professionnel·le·s de façon interactive durant le processus de traduction, qui faciliterait leur travail et soutiendrait la démarche créative, principalement si le système est entraîné par ceux-ci et celles-ci sur des traductions personnelles, comme nous le soutenons. Ce scénario implique toutefois de repenser l'interaction humain-machine, puisque la manière dont la TA est intégrée à la démarche — soit comme un dispositif de prétraitement qui prémâche l'ensemble travail, soit comme une aide simultanée au processus de traduction — représente un facteur déterminant de la qualité finale du texte cible.

De même que la complexité d'un texte peut être radicalement opposée pour la machine ou pour l'humain, la TA pourrait offrir des avantages complémentaires à la biotraduction. L'atout de ces systèmes automatiques, en effet, est qu'ils peuvent repérer et restituer des schémas autrement invisibles à l'œil nu. Ceux-ci seront donc plus pertinents s'ils sont issus de traductions personnelles, étant donné que des outils individualisés pourront alors intégrer de véritables choix de traduction et maintenir la cohésion lexicale ou syntaxique du texte, voire à un niveau plus abstrait encore, comme l'évoquent nos résultats initiaux. Ces échos de traduction n'ont dès lors pas vocation à couvrir la voix des traducteur·trice·s. Au contraire, ils renforcent à leur tour la cohérence stylistique de l'ouvrage dans son ensemble, ce qui peut être bienvenu si l'on tient compte de la masse de travail demandée par la traduction d'une œuvre ou d'une série littéraire. Ils offrent également une sorte de second regard qui peut appuyer et alimenter l'ingéniosité humaine, à l'instar de la TLAO, notamment en fournissant une ou plusieurs suggestions dont il est possible de s'inspirer pour résoudre un point de difficulté particulier et repérer d'éventuelles erreurs d'inattention ou d'interprétation, ou en accélérant la traduction et en dégageant du temps pour les passages plus créatifs.

À l'inverse, la TA amène de nouvelles interrogations concernant la propriété intellectuelle, le droit sur les données et la visibilité. Les risques n'ont néanmoins pas uniquement trait au statut des professionnel·le·s, mais aussi à la qualité et à la créativité des traductions produites. La majeure partie des préoccupations soulevées découlerait ainsi d'une utilisation aveugle et non raisonnée de cette technologie, en particulier si elle a uniquement vocation à remplacer l'humain pour des raisons de coût et de productivité [Taivalkoski-Shilov, 2019]. De la même manière, il sera pareillement important de voir quel serait l'impact sur la créativité. Est-ce que l'intervention de la TA conduirait à une réduction de cet aspect, comme l'indiquent les résultats de Guerberof-Arenas et Toral [2020], ou est-ce que le fait d'être familier avec la post-édition pourrait mitiger ce constat, comme le laissent penser les conclusions Nunes Vieira et al. [2020] ? Toutes ces analyses plus fines sur les aspects humains de la TLA sont impératives, mais elles sont rendues d'autant plus difficiles que la traduction est une tâche éminemment subjective, que les résultats s'en retrouvent bien souvent contradictoires et qu'il n'existe, à l'heure actuelle, probablement aucun juge familiarisé à la fois avec la littérature et la post-édition.

VI CONCLUSION

Dans cet article, nous avons soulevé la nécessité de poser un regard objectif sur la question de la traduction littéraire automatique et le besoin d'adapter des systèmes de TA dans ce sens, en les entraînant sur des jeux de données issus de ce même domaine. Nous avons répondu à cette observation par une expérience qui n'avait plus été renouvelée pour la paire anglais-français depuis les travaux de Besacier sur la TA probabiliste. À ce titre, nous avons mis au point un modèle de traduction automatique en comptant, d'une part, sur les architectures neuronales LSTM et *Transformer*, et, d'autre part, sur un corpus littéraire caractéristique du genre de la *fantasy* qui nous a permis d'affiner un système générique sur la production d'une auteure et d'une traductrice. Nous avons également vu, par comparaison avec d'autres essais, que différents systèmes testés sur diverses œuvres peuvent donner des résultats très variables. Ce qui ressort systématiquement de ces études et de la nôtre, cependant, ce sont les gains liés au passage de la traduction automatique neuronale vis-à-vis de la traduction probabiliste, et ceux de l'architecture *Transformer* en particulier. En affinant ces systèmes sur des textes littéraires, mais plus encore sur des données issues d'un genre spécifique et produites par une auteure et une traductrice déterminées, il nous a été possible d'atteindre en outre de bien meilleures performances, et ce, sur un roman visiblement et étonnement complexe pour la TA.

La littérature, notamment les ouvrages de fiction similaires à celui que nous mettons à l'essai dans cet article, continue de poser néanmoins des obstacles pour la TA, comme le laissent penser les scores toujours très bas de nos systèmes génériques et des services de traduction en ligne. Les scores obtenus par notre système adapté à la littérature, bien que très encourageants, demeurent aussi sans trop de surprise en deçà de ce que l'on peut observer dans d'autres domaines. Il nous faut tout de même rappeler que nous avons choisi et testé ici un cas singulièrement riche en défis, et que si des auteurs comme Toral et Way [2018] obtiennent des scores proches pour certains ouvrages avec leur modèle entraîné entièrement sur de la littérature, celui-ci parvient à réaliser de meilleures performances sur d'autres ouvrages. Le système utilisé dans notre démarche exploratoire reste d'ailleurs simple et nous imaginons que l'ajout de données supplémentaires ou que l'utilisation d'un système plus robuste comme il en existe beaucoup aujourd'hui pourrait améliorer encore les résultats.

Nous pouvons ainsi voir que même si la TA n'est pas près de remplacer l'humain et que le champ plus spécifique de la TLA reste encore un domaine de recherche très jeune, celle-ci demeure toutefois un outil intéressant, puisque notre expérience montre en effet qu'il est possible d'adapter un système non seulement à la littérature, mais aussi au style d'un·e traducteur·trice. Certes, ces développements technologiques nous éloignent immuablement de l'image traditionnelle de saint Jérôme, peignant à sa place un tableau où l'humain se trouve entouré d'outils très variés, sans même sans y compter la TA. Outils qui posent par ailleurs de nombreuses questions d'ordre éthique et sociétal que nous n'avons pas pu aborder ici, mais qu'il pourrait être intéressant selon nous de mettre sur la table avant l'arrivée de la TA en littérature, et qui pourraient peut-être concerner d'autres secteurs du métier dans lesquels elle est déjà bien présente. Dans la même optique, l'adaptation à ce domaine nous semble ouvrir la voie vers une autre manière d'envisager la TA, avec des outils adaptatifs qui refléteraient réellement le style des traducteur·trice·s. Ces traits personnels se reflètent nécessairement dans les traductions et peuvent influencer sa bonne réception par le destinataire, comme le soulignent Mirkin et al. [2015], mais ils nous semblent plus prégnants encore dans le cas de la littérature. Or, ces auteur·e·s notent que les modèles de TA généralistes qui existent actuellement effacent ces informations, tout comme Kenny et Winters [2020] attestent de l'effet neutralisant de ces mêmes systèmes génériques sur la voix des traducteur·trice·s. Des outils adaptés, en revanche, pourraient contribuer à renforcer ces éléments. Cela n'est pourtant possible que s'ils sont entraînés par les professionnel·le·s, sur leurs propres traductions, idéalement en combinaison avec d'autres aides telles que la traduction assistée par ordinateur ou d'autres logiciels d'exploration de corpus. À l'avenir, nous envisageons d'ailleurs justement de pousser plus avant l'évaluation des textes littéraires traduits par notre machine adaptée à la littérature, et nous projetons de tracer le parallèle entre l'application de la TA en littérature et sa mise à l'essai ou son usage actuel dans d'autres domaines culturels.

REMERCIEMENTS

Nous tenons à remercier Laurent Besacier, pour ses conseils et pour avoir contribué à la réalisation de ce projet.

Merci également aux relectrices pour leurs corrections et suggestions d'amélioration.

Les scripts et le matériel utilisés pour ce projet, à l'exception des données d'entraînement protégées par le droit d'auteur, seront mis en ligne à l'adresse suivante : https://gitlab.uliege.be/dhansen/literary-machine-translation.

Références

A. Araabi et C. Monz. Optimizing Transformer for Low-Resource Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelone, 2020. ICCL. doi : 10.18653/v1/2020.coling-main.304.

D. Bahdanau, K. Cho, et Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In Y. Bengio et Y. LeCun, éditeurs, *3rd International Conference on Learning Representations : Conference Track Proceedings*, pages 1–15, San Diego (CA), 2014. arXiv : 1409.0473.

L. Bentivogli, A. Bisazza, M. Cettolo, et M. Federico. Neural versus Phrase-Based Machine Translation Quality : a Case Study. In J. Su, K. Duh, et X. Carreras, éditeurs, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin (TX), 2016. ACL. doi : 10.18653/v1/D16-1025.

L. Besacier. Traduction automatisée d'une oeuvre littéraire : une étude pilote. In P. Blache, F. Béchet, et B. Bigi, éditeurs, *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, pages 389–394, Marseille, 2014. ATALA. URL : http://talnarchives.atala.org/TALN/TALN-2014/taln-2014-court-001.pdf.

R. Blin. Neural machine translation, corpus and frugality. *ArXiv* preprint, pages 1–7, 2021. arXiv : 2101.10650.

L. Bowker et J. B. Ciro. *Machine Translation and Global Research : Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Group Publishing, Bingley, 2019.

J. Cambreleng. L'observatoire de la traduction automatique. *La traduction littéraire et SHS à la rencontre des technologies de la traduction : enjeux, pratiques et perspectives*, Université Toulouse-Jean Jaurès, France, 2020.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, 2014. ACL. doi : 10.3115/v1/D14-1179.

C. Chu et R. Wang. A Survey of Domain Adaptation for Neural Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe (NM), 2018. ACL. URL : https://aclanthology.org/C18-1111.pdf.

N. Froeliger. *Les Noces de l'analogique et du numérique*. Presses Universitaires du Septentrion, Villeneuve d'Ascq, 2013.

M. Ghazvininejad, Y. Choi, et K. Knight. Neural Poetry Translation. In M. Walker, H. Ji, et A. Stent, éditeurs, *Proceedings of NAACL-HLT 2018*, volume 2, pages 67–71, New Orleans (LO), 2018. ACL. doi : 10.18653/v1/N18-2011.

A. Guerberof-Arenas et A. Toral. The Impact of Post-Editing and Machine Translation on Creativity and Reading Experience. *Translation Spaces*, 9(2) : 255–282, 2020. doi : 10.1075/ts.20035.gue.

D. Hansen. Les lettres et la machine : un état de l'art en traduction littéraire automatique. In P. Denis, N. Grabar, A. Fraisse, R. Cardon, B. Jacquemin, E. Kergosien, et A. Balvet, éditeurs, *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, pages 28–45, Lille, 2021. ATALA. HAL : hal-03265904.

D. Hansen. The Figure of the Literary Translator Amid New Technologies, à paraître.

D. Hansen et P.-Y. Houlmont. A Snapshot into the Possibility of Video Game Machine Translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas*, volume 2, pages 257–269, Orlando (FL), 2022. AMTA. URL : https://aclanthology.org/2022.amta-upg.18.pdf.

W. J. Hutchins. The Georgetown-IBM Experiment Demonstrated in January 1954. In R. E. Frederking et K. B. Taylor, éditeurs, *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas : Technical Papers*, pages 102–114, Washington DC, 2004. Springer. URL : https://aclanthology.org/www.mt-archive.info/00/AMTA-2004-Hutchins.pdf.

D. Kenny et M. Winters. Machine translation, ethics and the literary translator's voice. *Translation Spaces*, 9(1) : 123–149, 2020. doi : 10.1075/ts.00024.ken.

G. Klein, Y. Kim, Y. Deng, J. Senellart, et A. Rush. OpenNMT : Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, 2017. ACL. URL : https://aclanthology.org/P17-4012.pdf.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, et E. Herbst. Moses : Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, 2007. ACL. URL : https://aclanthology.org/P07-2045.pdf.

T. Kudo. Subword Regularization : Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 66–75, Melbourne, 2018. ACL. doi : 10.18653/v1/P18-1007.

T. Kuzman, Š. Vintar, et M. Arčan. Neural Machine Translation of Literary Texts from English to Slovene. In J. Hadley, M. Popović, H. Afli, et A. Way, éditeurs, *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, Dublin, 2019. EAMT. URL : https://www.aclweb.org/anthology/W19-7301.pdf.

F. Landragin. *Les machines à langage dans la science-fiction*. Le Bélial', Moret-Loing-et-Orvanne, 2020.

H. Lee-Jahnke. Le traducteur, passeur entre les cultures. In M. Forstner et H. Lee-Jahnke, éditeurs, *Regards sur les aspects culturels de la communication*, pages 61–86. Peter Lang, Berne, 2005.

R. Loock. La plus-value de la biotraduction face à la machine : Le nouveau défi des formations aux métiers de la traduction. *Traduire*, 241 : 54–65, 2019. doi : 10.4000/traduire.1848.

R. Loock. Introduction de la journée d'études TQ2020. *Traduction et Qualité 2020 : Biotraduction et traduction automatique*, Université de Lille, France, 2020. URL : https://tq2020.sciencesconf.org/.

M. Loponen. Translating Irrealia : Creating a Semiotic Framework for the Translation of Fictional Cultures. *Chinese semiotic studies*, 2(1) :165–175, 2009. doi : 10.1515/css-2009-0117.

M.-T. Luong, H. Pham, et C. D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In L. Màrquez, C. Callison-Burch, et J. Su, éditeurs, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbonne, 2015. ACL. doi : 10.18653/v1/D15-1166.

E. Matusov. The Challenges of Using Neural Machine Translation for Literature. In J. Hadley, M. Popović, H. Afli, et A. Way, éditeurs, *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, 2019. EAMT. URL : https://www.aclweb.org/anthology/W19-7302.pdf.

S. Mirkin et J.-L. Meunier. Personalized Machine Translation : Predicting Translational Preferences. In L. Màrquez, C. Callison-Burch, et J. Su, éditeurs, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbonne, 2015. ACL. doi : 10.18653/v1/D15-1238.

S. Mirkin, S. Nowson, C. Brun, et J. Perez. Motivating Personality-aware Machine Translation. In L. Màrquez, C. Callison-Burch, et J. Su, éditeurs, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbonne, 2015. ACL. doi : 10.18653/v1/D15-1130.

J. Moorkens, A. Toral, S. Castilho, et Way A. Translators' Perceptions of Literary Post-editing using Statistical and Neural Machine Translation. *Translation Spaces*, 7(2) : 240–262, 2018. doi : 10.1075/ts.18014.moo.

Y. Moslem. Domain Adaptation Techniques for Neural Machine Translation. *AMTA 2020*, 2020.

M. Müller, A. Rios, et R. Sennrich. Domain robustness in neural machine translation. In M. Denkowski et C. Federmann, éditeurs, *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*, volume 1, pages 151–164. AMTA, 2020. URL : https://aclanthology.org/2020.amta-research.14.pdf.

L. Nunes Vieira, X. Zhang, R. Youdale, et M. Carl. Machine Translation and Literary Texts : A Network of Possibilities. *Creative Translation and Technologies Expert Meeting*, Université de Surrey, Royaume-Uni, 2020.

F. J. Och et H. Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1) : 19–51, 2003. doi : 10.1162/089120103321337421.

K. Papineni, S. Roukos, T. Ward, et W.-J. Zhu. BLEU : A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia (PA), 2002. ACL. doi : 10.3115/1073083.1073135.

M. Post. A Call for Clarity in Reporting BLEU Scores. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, et K. Verspoor, éditeurs, *Proceedings of the Third Conference on Machine Translation : Research Papers*, pages 186–191, Bruxelles, 2018. ACL. doi : 10.18653/v1/W18-6319.

D. Rao. GPT-3 and a Typology of Hype. *Page Street Labs*, 2020. URL : https://pagestlabs.substack.com/p/gpt-3-and-a-typology-of-hype.

C. Rossi. Les usages actuels de la traduction automatique. *Atelier DigitHum 2019*, ENS, France, 2019. URL : https://digithum.huma-num.fr/atelier/2019/4_rossi.php.

R. Sennrich et B. Zhang. Revisiting Low-Resource Neural Machine Translation : A Case Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, 2019. ACL. doi : 10.18653/v1/P19-1021.

R. Sennrich, B. Haddow, et A. Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 86–96, Berlin, 2016. ACL. doi : 10.18653/v1/P16-1009.

A. Stolcke. SRILM – An extensible language modeling toolkit. In J. H. L. Hansen et B. L. Pellom, éditeurs, *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP'2002)*, pages 901–904, Denver (CO), 2002. ISCA.

I. Sutskever, O. Vinyals, et Q. V. Le. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, et K. Q. Weinberger, éditeurs, *NIPS'14 : Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2, pages 3104–3112, Montréal, 2014. MIT Press. URL : https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

K. Taivalkoski-Shilov. Ethical Issues Regarding Machine(-Assisted) Translation of Literary Texts. *Perspectives : Studies in Translation Theory and Practice*, 27(5) : 689–703, 2019. doi : 10.1080/0907676X.2018.1520907.

J. Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, 2012. ELRA. URL : http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

A. Toral et A. Way. Translating Literary Text between Related Languages using SMT. In A. Feldman, A. Kazantseva, S. Szpakowicz, et C. Koolen, éditeurs, *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 123–132, Denver (CO), 2015a. ACL. doi : 10.3115/v1/W15-0714.

A. Toral et A. Way. Machine-assisted translation of literary text : A case study. *Translation Spaces*, 4(2) : 241–268, 2015b. doi : 10.1075/ts.4.2.04tor.

A. Toral et A. Way. What Level of Quality can Neural Machine Translation Attain on Literary Text ? In J. Moorkens, S. Castilho, F. Gaspari, et S. Doherty, éditeurs, *Translation Quality Assessment : From Principles to Practice*, pages 263–287. Springer, New York (NY), 2018. arXiv : 1801.04962.

T. Van de Cruys. La génération automatique de poésie en français. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019*, volume 1, pages 113–126. ATALA, 2019. URL : https://aclanthology.org/2019.jeptalnrecital-long.8.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, et I. Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et R. Garnett, éditeurs, *NIPS'17 : Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach (CA), 2017. Curran Associates. URL : https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

A. Zydroń. De-Demonizing AI Regarding Localization. In S. Chambers, J. Esteves-Ferreira, J. M. Macan, J. Moorkens, R. Mitkov, M. Recort Ruiz, O.-M. Stefanov, et J.-M. Vande Walle, éditeurs, *Translating and the Computer 41*, pages 137–144, Londres, 2019. Éditions Tradulex. URL : https://www.asling.org/tc41/wp-content/uploads/TC41-Proceedings_137-144.pdf.