Enseigner la compilation et l'exploitation de corpus monolingues spécialisés pour la traduction : retour sur expériences et suggestions

Rudy Loock

Enseigner la compilation et l'exploitation de corpus monolingues spécialisés pour la traduction : retour sur expériences et suggestions

Rudy Loock, Université de Lille et UMR 8163 Savoirs, Textes, Langage du CNRS

Dans cet article, nous dressons un bilan critique de la formation à la compilation et à l'exploitation de corpus électroniques (que nous souhaitons renommer « bases de données linguistiques ») pour la traduction spécialisée, en nous concentrant particulièrement sur les corpus monolingues en langue spécialisée (domaine, mais aussi genre textuel), compilés pour la circonstance et exploités avec un concordancier dans le contexte d'un projet de traduction (corpus dits « DIY » pour « *Do It Yourself* »). À partir des résultats d'enquêtes de terrain relatives aux compétences acquises et à l'utilisation des outils de corpus dans le cadre de la vie professionnelle, et du constat que cette utilisation est loin d'être généralisée, nous formulons un certain nombre de propositions afin d'optimiser ce type d'enseignement avec notamment un repositionnement stratégique dans la façon dont les corpus électroniques sont présentés dans les formations universitaires aux métiers de la traduction. L'ensemble de ces réflexions s'appuie sur une expérience de dix années d'un enseignement consacré aux outils de corpus (compilation, exploitation, analyse) au sein d'une formation universitaire à la traduction spécialisée.

Mots-clefs : corpus, base de données linguistiques, corpus spécialisé, traduction, formation

In this article, we present a critical review of training in the compilation and use of electronic corpora (which would like to rename as "linguistic databases") for specialised translation, with a particular focus on monolingual corpora for a specialised variety of language (domain, textual genre) that are compiled specifically and used with a concordancer for a translation project (so-called DIY corpora for "Do It Yourself"). After presenting the results of surveys on post-training corpus-related skills and the use of corpus tools in the context of professional life, and following the observation that such use is far from widespread, we are making a number of suggestions in order to improve this type of teaching. Among these is a strategic shift in the way electronic corpora are introduced in university translation programmes. Our suggestions are inspired by a ten-year experience in teaching the use of corpus tools (compilation, exploitation, analysis) in a specialised translation university course.

Keywords: corpus, linguistics database, specialized corpus, translation, training

1. L'utilisation des corpus spécialisés en traduction : définitions, compétences, et étude de cas

1.1. Définitions

Parmi les nombreux outils à disposition des spécialistes de la traduction spécialisée figurent les corpus électroniques, soit des ensembles de textes correspondant à des échantillons représentatifs d'une langue ou d'un type de discours, à exploiter de façon automatique. Il peut s'agir de bases de données disponibles en ligne, p. ex. le site english-corpora (https://www.english-corpora.org/) pour la langue anglaise ou le portail Sketch Engine (https://www.Sketch Engine.eu/) et ses plus de 800 corpus pour une centaine de langues, mais il peut également s'agir de corpus compilés pour la circonstance, avec pour objectifs possibles en traduction une meilleure compréhension du texte source, une amélioration de la qualité rédactionnelle en langue cible, ou encore la compilation de glossaires. Dans ce second cas, les professionnels de la traduction peuvent notamment compiler et exploiter des bases de données réunissant des textes monolingues spécialisés en langue cible. Ce type de corpus dits « DIY » (pour « *Do-It-Yourself* », voir p. ex. Sánchez-Gijón, 2009 ; Scott, 2012 ; Loock 2016a), compilés pour la circonstance manuellement ou de façon semi-automatisée, permettent de rassembler des textes portant sur une thématique donnée et relevant d'un genre spécifique (article scientifique, rapport d'une organisation internationale, mode d'emploi…). Ces corpus DIY sont la plupart du temps exploités hors ligne à l'aide d'un concordancier, même s'il est possible de compiler et d'exploiter ses propres corpus en ligne à partir de ses propres données, via le portail Sketch Engine par exemple.

Afin de produire une traduction de qualité répondant aux normes d'un domaine et d'un genre spécialisés, la recherche d'informations linguistiques sur la langue cible (terminologie, phraséologie, grammaire, organisation du discours) est absolument nécessaire, et les outils traditionnels comme les dictionnaires et glossaires en ligne ne suffisent pas toujours (voir *infra*). Les corpus de langue générale consultables en ligne montrent également leurs limites dans ce type de configuration puisque par définition la langue qui y est représentée ne concerne pas un domaine ou un genre précis. Quant aux recherches internet par le biais d'un moteur de recherche, elles posent des problèmes de qualité et de fiabilité des informations, problèmes exacerbés ces dernières années par l'omniprésence de textes générés ou traduits de façon automatique, optimisés pour le référencement naturel, et de qualité hautement discutable (Thompson, Dhaliwal, Frisch, Domhan & Federico, à paraître ; Bevendorff, Wiegmann, Potthast & Stein, 2024). C'est ainsi que les professionnels peuvent être amenés à compiler leurs propres corpus DIY monolingues spécialisés en langue cible, dans notre cas le français, à partir de textes collectés sur l'internet, par exemple sur le site EUR-Lex pour les textes juridiques, le site du Fonds monétaire international pour les textes économiques, ou encore les sites de l'Organisation mondiale de la santé ou de l'Agence régionale de santé pour les documents médicaux. Ces bases de données peuvent contenir de plusieurs centaines à plusieurs dizaines de milliers de mots, être compilées manuellement ou de façon semi-automatique

via un logiciel dédié (p. ex. BootCaT[1]), et être exploitées hors ligne grâce à un concordancier, p. ex. AntConc, LancsBox, TextStat, ou encore Wordsmith Tools[2]. L'objectif est alors de consulter la langue spécialisée telle qu'elle est utilisée par les experts d'un domaine donné afin de s'en inspirer et d'assurer l'« invisibilité » demandée aujourd'hui sur le marché, où l'on attend des professionnels qu'ils fournissent des textes traduits ne laissant pas transparaître qu'ils sont le résultat d'une traduction.

1.2. Quelles compétences ?

Parmi les compétences nécessaires aux futurs professionnels qui souhaitent exercer dans le secteur de la traduction spécialisée, les compétences relatives aux nouvelles technologies sont désormais incontournables. Ainsi, le référentiel de compétences du réseau *European Master's in Translation* (EMT) de la Direction générale de la traduction de la Commission européenne[3] considère ces compétences comme l'un des 5 piliers de compétences à maîtriser à l'issue d'une formation aux métiers de la traduction. Parmi les 36 compétences du référentiel, 6 concernent la maîtrise des technologies, et on peut considérer que 4 d'entre elles (voir définitions en (1)) sont relatives à l'utilisation de bases de données linguistiques/corpus électroniques.

(1) Compétence 15 : utiliser les applications informatiques les plus pertinentes, y compris l'éventail complet des logiciels bureautiques, et s'adapter rapidement aux nouveaux outils et ressources informatiques après avoir évalué de manière critique leur pertinence et l'incidence du changement sur leurs pratiques de travail

Compétence 16 : utiliser efficacement les moteurs de recherche et les outils de corpus, d'analyse de texte, de traduction assistée par ordinateur (TAO) et d'assurance qualité (AQ), le cas échéant

Compétence 18 : comprendre les bases des systèmes de traduction automatique et leur incidence sur le processus de traduction, et intégrer la traduction automatique dans un flux de travail de traduction, le cas échéant

Compétence 19 : reconnaître l'importance et la valeur des données de traduction et des données linguistiques (preuve d'une éducation aux données)

Les compétences 15 et 16 sont des compétences directes en matière d'utilisation pertinente et efficace, tandis que les compétences 18 et 19 sont indirectes puisque la compilation ou l'exploitation d'un corpus requièrent de comprendre le rôle crucial des données. Une formation à ce que sont les données et comment les sélectionner est alors indispensable. L'utilisation raisonnée des outils de traduction automatique, par exemple, implique une analyse pertinente des sorties de TA avant post-

[1] https://bootcat.dipintra.it/

[2] Ces différents concordanciers sont disponibles aux adresses suivantes : https://www.laurenceanthony.net/software/antconc/, http://corpora.lancs.ac.uk/lancsbox/, https://neon.niederlandistik.fu-berlin.de/en/textstat/, https://www.lexically.net/wordsmith/.

[3] https://commission.europa.eu/system/files/2023-01/emt_competence_fwk_2022_fr.pdf

édition et une conscience des biais induits par les données qui alimentent le moteur (biais algorithmiques, mais aussi sexistes et discriminatoires).

De nombreuses formations aux métiers de la traduction proposent aujourd'hui un ou des enseignements relatifs à la compilation et à l'exploitation de données rassemblées en corpus. À partir de l'analyse de différentes enquêtes menées aux niveaux national français et européen, Frérot & Karagouch (2016) avancent ainsi que de tels enseignements sont proposés par près de deux formations sur trois, avec toutefois une multiplicité d'approches qui ne mettent pas toujours en avant l'utilisation des corpus comme outils d'aide à la traduction. Une étude plus récente concernant l'intégration d'enseignements relatifs aux corpus dans les formations universitaires à l'échelle internationale (Mikhailov, 2022) témoigne d'une pratique bien plus marquée, avec 72 formations sur les 91 interrogées ayant mis en place ce type d'enseignements. L'état des lieux est toutefois biaisé, comme le reconnaît l'auteur lui-même, dans la mesure où l'enquête, ouverte à tous mais concernant directement cette question, aura sans doute reçu des réponses majoritairement de la part de formations ayant intégré les corpus électroniques dans la formation de leurs étudiants. L'étude reste néanmoins informative dans la mesure où elle fournit un instantané des pratiques (types de corpus abordés, avec quels objectifs). Cette progression dans les formations universitaires s'explique par une amélioration significative de l'accessibilité mais aussi de l'ergonomie des outils de corpus, qui ont beaucoup progressé ces dernières années. Également, la présence de données linguistiques dans de nombreux outils d'aide à la traduction, et notamment les outils de traduction en ligne, peut venir expliquer cette évolution, puisque plus que jamais, une formation aux données est importante. L'arrivée des outils d'intelligence artificielle que sont les grands modèles de langage (*large language models*) ne pourra que venir renforcer ce besoin : les récents outils d'IA génératives s'appuient en effet sur de grands ensembles de textes, et les questions relatives à la qualité, la fiabilité, la représentativité, ou encore les biais, ne sont pas sans rappeler les enjeux bien connus posés par les outils de corpus.

Toute une littérature a par ailleurs fait la démonstration de la pertinence et de l'efficacité du recours aux corpus électroniques spécialisés comme outils d'aide à la traduction, à la révision, ou encore plus récemment à la post-édition de traduction automatique (voir p. ex. Bowker & Pearson, 2002 ; Kübler, 2003 ; Zanettin, Bernardini & Stewart, 2003 ; Frankenberg-Garcia, 2015 ; Loock, 2016b ; Kübler, Mestivier & Pecman, 2022 ; Giampieri, 2021). Nous ne proposons pas une synthèse de ces travaux ici, mais proposons dans la section qui suit d'illustrer l'utilisation possible des corpus DIY monolingues comme outils d'aide à la traduction par le biais d'un exercice effectué avec nos étudiants.

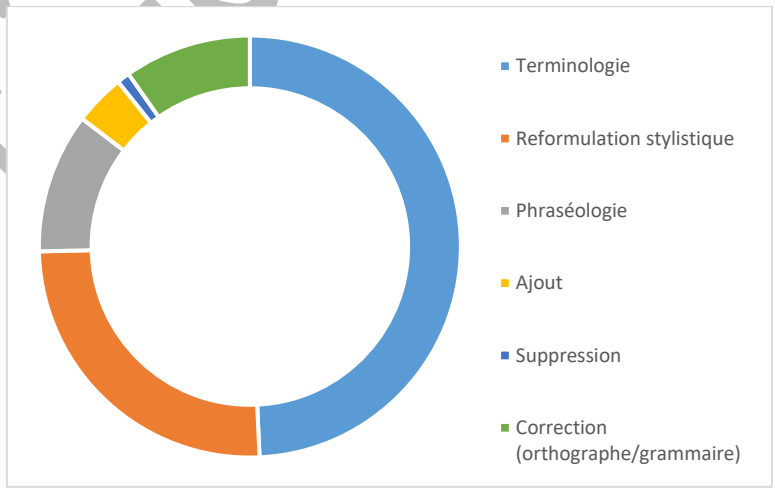
1.3. Illustration : l'exemple des plugins pour ChatGPT

Afin d'illustrer l'affirmation formulée en fin de section précédente, nous revenons ici sur un exercice effectué en novembre 2023 avec les étudiants inscrits en première année du programme de master « Traduction spécialisée multilingue » (TSM) de l'Université de Lille. Dans le cadre de deux enseignements pour lesquels nous pratiquons le décloisonnement[4], à savoir un cours de traduction

[4] Le décloisonnement pédagogique est une approche éducative qui vise à transcender les limites traditionnelles des enseignements pour développer des compétences données ou effectuer une tâche précise.

de l'anglais vers le français avec pour spécialisation les technologies de l'information et un cours consacré aux outils de corpus, nous avons demandé aux étudiants de traduire en français un document de 396 mots relatif aux plugins pour ChatGPT, directement extrait du site internet de la société OpenAI (https://openai.com/blog/chatgpt-plugins). Pour cette première traduction à préparer de façon autonome chez eux, les étudiants (n=27) avaient la possibilité d'utiliser tous les outils à leur disposition, y compris la traduction automatique. Une fois cette première traduction livrée, ils ont reçu une formation à la compilation de corpus DIY et à l'exploitation via un concordancier, en l'occurrence AntConc (Anthony, 2023), en lien avec la thématique des plugins pour ChatGPT. Des documents rédigés directement en français sur le sujet et provenant de sites informatiques spécialisés ont été rassemblés pour former un corpus de 17 519 tokens provenant de 8 sources différentes. Armés de ce corpus DIY, les étudiants ont été invités à réviser leur première traduction, en apportant ou non (il n'y avait pas de caractère obligatoire) des modifications à leur traduction initiale.

Le premier constat est que la majeure partie des étudiants a souhaité effectuer des changements : seuls 3 d'entre eux n'ont pas souhaité modifier leur traduction. En moyenne, les autres étudiants ont effectué 8,5 modifications. Une comparaison automatique[5] de la traduction initiale et de la traduction révisée montre un taux de différence moyen de 6,44 %, ce qui cache toutefois une disparité puisque le taux de différence s'étale de 0,97 % à 18,35 % (valeur médiane = 5,12 %). La majeure partie des révisions concerne la terminologie (49,2 %) et notamment la terminologie spécialisée : ainsi le mot *extension* a souvent été remplacé par *plugin*, les termes *période alpha* ou *période d'essai* par *phase alpha*, le terme *modèle linguistique* remplacé par *modèle de langage*. Le deuxième type de modifications le plus fréquent (25,4 %) a porté sur les reformulations en vue d'améliorer le style en prenant en compte les questions de registre et collocationnelles (*utilisations possibles > usages potentiels, et > ou encore, réussir notre objectif > accomplir notre mission, interaction entre l'homme et l'IA > interaction humaine avec l'IA*). D'autres révisions ont concerné la phraséologie, notamment l'emploi des prépositions (10,7 %), la correction orthographique ou grammaticale (9,7 %), ainsi que quelques ajouts (3,9 %) ou suppressions (0,9 %). La Figure 1 schématise ces résultats.



[5] Le taux de différence entre traductions initiales et traductions révisées a été calculé automatiquement grâce à l'outil 'Compare texts' disponible à l'adresse suivante : https://countwordsfree.com/comparetexts.

Figure 1. Types de révisions effectuées par les étudiants après exploitation du corpus

Naturellement, toutes les modifications apportées n'ont pas toujours été ni utiles ni correctes comme cela a pu être discuté avec les étudiants lors de la traduction en cours de ce texte ; de la même manière, certaines révisions, notamment les corrections d'erreurs d'orthographe ou de grammaire ainsi que les améliorations stylistiques, ne sont pas une influence directe de l'utilisation du corpus. On voit bien en revanche que l'observation de données linguistiques rédigées par des spécialistes sur un sujet donné entraîne un certain nombre de révisions de la part des étudiants. Par manque de place, nous n'entrons pas davantage dans le détail ici.

2. Défis et enjeux

2.1. Un outil d'aide à la traduction parmi tant d'autres

Les traducteurs professionnels ont aujourd'hui accès à une multitude d'outils et de technologies : traduction assistée par ordinateur (TAO), traduction automatique (TA), outils générateurs de texte s'appuyant sur l'intelligence artificielle (ChatGPT étant sans doute le plus connu d'entre eux au moment où nous rédigeons ces lignes), logiciels de correction orthographique et grammaticale, outils d'assurance qualité, sans oublier les incontournables dictionnaires et glossaires en ligne en tous genres parmi lesquels les apprentis traducteurs doivent apprendre à naviguer (voir Rothwell, Moorkens, Fernández-Parra, Drugan & Austermuehl, 2023 pour un état des lieux récents). La majeure partie de ces outils s'appuie sur des bases de données linguistiques ou corpus : ainsi, les mémoires de traduction qui alimentent les logiciels de TAO ne sont rien d'autre que des corpus parallèles bilingues alignés au niveau de la phrase, les outils de TA et les nouveaux outils conversationnels s'appuyant sur l'intelligence artificielle exploitent de grands ensembles de données linguistiques, et nombreux sont les dictionnaires en ligne, monolingues ou bilingues, qui fournissent aujourd'hui des exemples issus de bases de données, qu'il s'agisse de l'internet ou d'un corpus spécifique, p. ex. https://www.collinsdictionary.com ou https://www.linguee.com/.

Au-delà, les bases de données linguistiques peuvent être exploitées non pas pour permettre le fonctionnement d'un outil, mais en tant que telles. Ainsi les traducteurs peuvent avoir recours à des corpus en ligne ou bien compiler leurs propres bases de données linguistiques à exploiter à l'aide d'un concordancier (voir définition *supra*), qui ne sont alors qu'un outil parmi d'autres, et leur maîtrise (compilation et exploitation dans le cas des corpus DIY) peut s'avérer plus fastidieuse que celle d'autres outils. Un suivi régulier de la promotion d'étudiants inscrits en 2023-2024 en première année dans la formation TSM témoigne en effet de cette appropriation lente et difficile, contrairement à d'autres outils pourtant découverts au même moment, à savoir au cours du premier

semestre de formation.[6] La Figure 2 montre ainsi l'évolution des outils utilisés par ces étudiants (n=24 à 28) entre le début de formation (septembre 2023) et la fin des enseignements avant le départ en stage (mars 2024) : si les étudiants semblent s'approprier très vite des outils comme Sketch Engine[7], qui propose toute une série de fonctionnalités (concordances, synonymes, profil linguistique d'un terme…) avec une utilisation par plus de 1 étudiant sur 2 dès le mois de novembre pour atteindre une proportion de 80 % au cours de l'année, on constate que l'utilisation des corpus DIY monolingues en langue spécialisée demeure extrêmement minoritaire, avec un maximum de 2 étudiants sur 10. À l'inverse, le recours à un logiciel de TAO, pourtant tout aussi technique, semble se faire de façon plus rapide.

[6] Ce suivi s'est fait tout au long de l'année universitaire dans le cadre d'un enseignement de la pratique de la traduction anglais-français, au sein duquel les étudiants sont libres d'utiliser l'ensemble des outils à leur disposition, qu'il s'agisse de préparations à faire à la maison ou d'examens sur site en salle informatique.

[7] Les étudiants de l'Université de Lille bénéficient d'un abonnement institutionnel à cet outil.

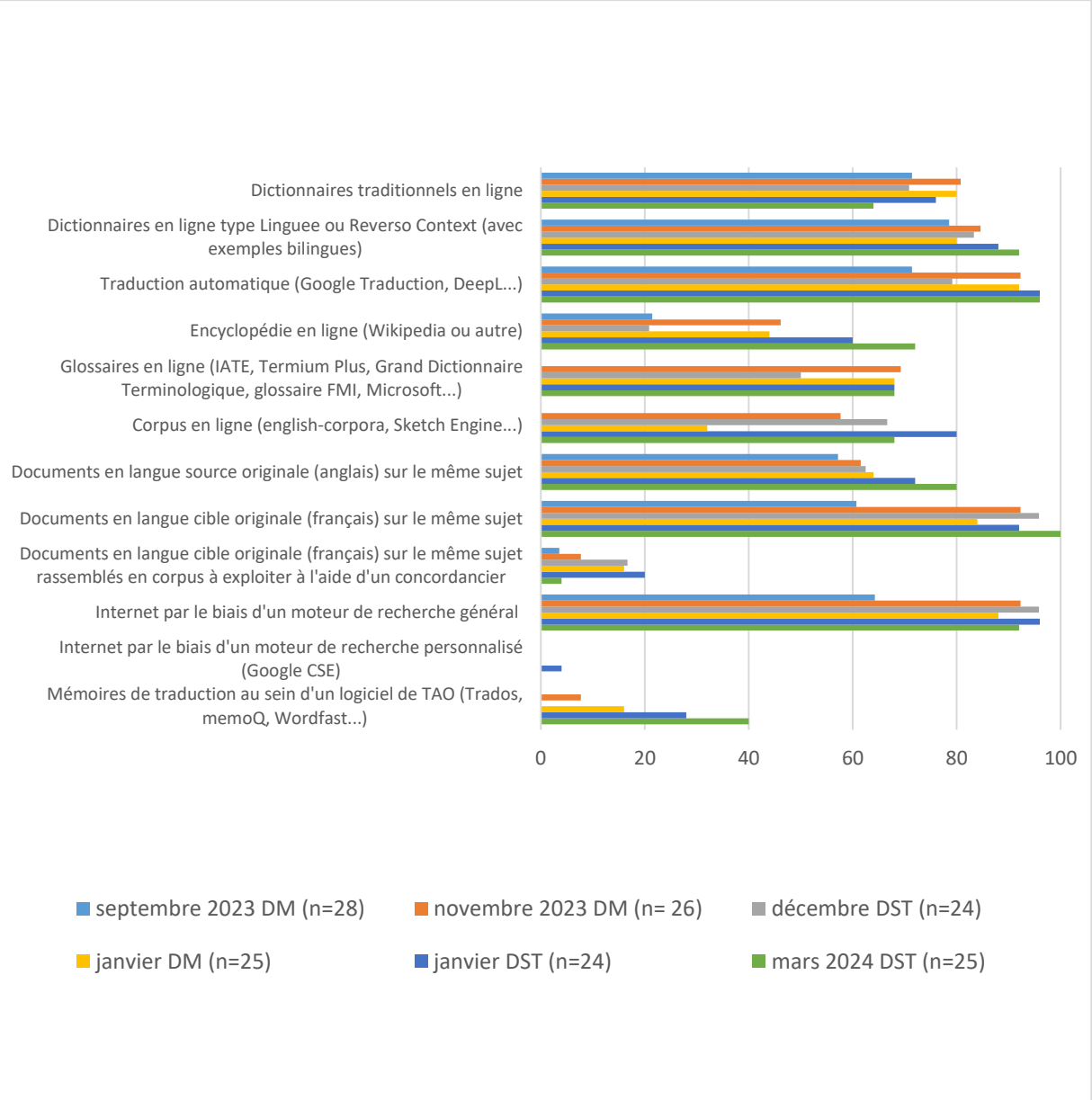


Figure 2. Fréquence d'utilisation (en %) des outils par les étudiants en première année de master pour les travaux effectués à la maison (DM) ou en situation d'examen (DST)

2.2. Des compétences durables mais pas toujours exploitées

Face à ces difficultés d'appropriation constatées, il nous a semblé pertinent de dresser un bilan critique de l'enseignement consacré aux outils de corpus (compilation, exploitation, analyse) au sein de la formation universitaire. Cet enseignement, d'une durée de 10 heures en première année de master et de 15 heures en seconde année, forme à l'utilisation de corpus en ligne et de corpus monolingues DIY exploités hors ligne à l'aide d'un concordancier. Il s'agit de fournir aux étudiants des outils visant à affiner la compréhension du texte source en langue étrangère, à améliorer la qualité du rédactionnel en langue cible, à évaluer le degré de formalisme et de technicité d'un document, à

comparer deux systèmes linguistiques, à effecteur de l'extraction terminologique pour compiler des glossaires, ou encore à analyser les textes traduits, y compris de façon automatique, afin de mettre au jour les spécificités linguistiques de la langue traduite qui la distingue de la langue originale. L'ensemble de cette approche, qui exploite les corpus comme outils d'aide à la traduction et comme outils de recherche en traductologie, relève de la traductologie de corpus (Loock, 2016b), ou *corpus-based translation studies* en anglais.

Afin de mesurer ce qu'il restait de cet enseignement chez les professionnels aujourd'hui en exercice, nous avons soumis un questionnaire à celles et ceux ayant obtenu leur diplôme entre 2014 et 2022. Il s'agissait de les interroger sur leur perception vis-à-vis de l'acquisition de compétences en matière de corpus à l'issue de la formation et aujourd'hui, soit n années après (n=1 à 9), mais aussi sur la mise en œuvre de ces compétences et leur utilisation en tant qu'outils d'aide à la traduction dans leur pratique professionnelle. Le questionnaire, disponible en annexe, contenait 14 questions et a été transmis début 2023 à 152 anciens étudiants de la formation de master « Traduction spécialisée multilingue » de l'Université de Lille par voie électronique. Nous avons reçu 42 réponses, soit un taux de réponse de 27,6 %. Le profil des répondants est le suivant : il s'agit d'anciens étudiants exerçant aujourd'hui dans le secteur de la traduction et appartenant de façon plutôt équilibrée aux différentes promotions avec une moyenne de 4,6 répondants par promotion (valeur minimale = 2, valeur maximale = 8 pour une valeur médiane de 4) avec pour principales paires de langues de travail anglais-français (41), espagnol-français (8), allemand-français (8), italien-français (5), suédois-français (4). La grande majorité d'entre eux (35, soit 83 %) exercent en tant qu'indépendants tandis que 5 (11,9 %) travaillent dans une agence de traduction ou une entreprise privée et 1 est fonctionnaire dans une organisation internationale. Deux tiers d'entre eux (28) ont une ou plusieurs spécialisations, parmi lesquelles la cybersécurité, le juridique, le tourisme, le médical, l'automobile, les loisirs créatifs, ou encore les jeux vidéo.

Ce qui ressort de cette enquête, c'est une perte progressive mais toutefois limitée des différentes compétences acquises lors de la formation universitaire (voir section 2.3 pour un propos plus nuancé, en fonction du statut). Ainsi, après avoir suivi les enseignements, les répondants s'estimaient capables d'utiliser les corpus en ligne pour améliorer la compréhension d'un texte source (76,2 %) ou la qualité de leur rédactionnel en langue cible (95,2 %), ou encore capables de compiler et d'exploiter des corpus hors ligne pour effectuer les mêmes tâches selon les mêmes proportions mais aussi pour compiler un glossaire bilingue (59 %). Ces résultats sont en recul lorsqu'il leur est demandé d'auto-évaluer leurs compétences aujourd'hui, soit entre 1 et 9 années après l'obtention du diplôme : les compétences relatives aux corpus en ligne ne sont plus maîtrisées pour les tâches mentionnées *supra* que par 64,3 % et 81 %, et celles relatives aux corpus DIY, qui sont l'objet de notre étude ici, que par 52,4 %, 59,5 % et 33,3 % respectivement. Au final, c'est l'utilisation des corpus DIY monolingues spécialisés qui souffre le plus d'une perte de compétences.

On pourrait malgré tout se satisfaire d'un tel résultat général, la perte de compétences étant estimée assez limitée, mais lorsque les répondants sont interrogés sur leur utilisation des outils de corpus dans le cadre professionnel aujourd'hui, les résultats sont moins encourageants. Ainsi, seul 1 répondant sur 3 utilise les corpus en ligne pour améliorer la compréhension du texte source et 1 sur 2 pour l'amélioration du texte cible. S'agissant des corpus DIY qui nous intéressent particulièrement ici, ils ne sont que 5 %, 3 %, et 7,2 % à les exploiter pour les 3 tâches mentionnées *supra*. Cela signifie que même pour les cas où les compétences sont perçues comme étant toujours maîtrisées n années après l'obtention du diplôme, elles ne sont pas nécessairement mises en pratique. Malgré tout,

83,3 % des répondants (35/42) restent convaincus que les corpus sont des outils d'aide à la traduction.

L'ensemble de ces résultats (compétences estimées acquises à l'issue de la formation, n années plus tard, et utilisation) est schématisé par la Figure 3.

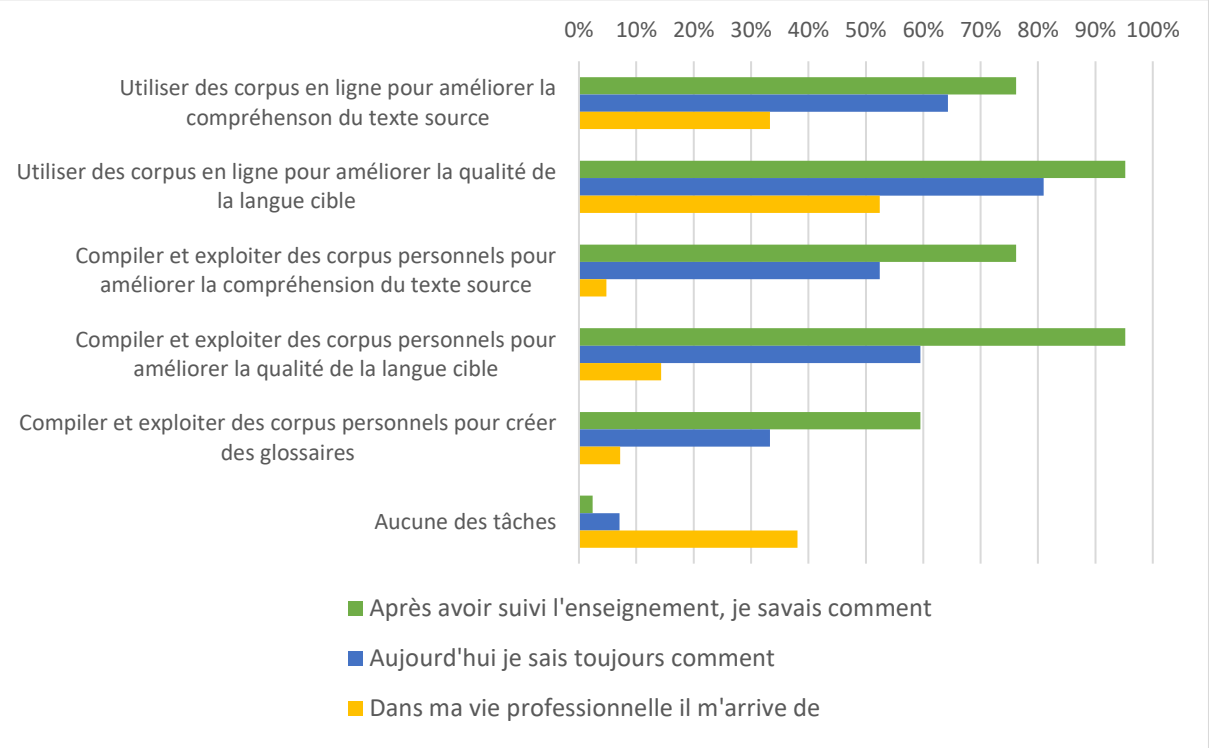


Figure 3. Compétences estimées acquises à l'issue de la formation et en 2023 ; utilisation dans le cadre professionnel

Interrogés sur leur (non) motivation à utiliser des corpus au moyen de réponses rédigées de façon libre, les répondants évoquent à la fois l'utilité des outils existants (2) et les problèmes d'accessibilité et de temps qui freinent leur utilisation (3), s'agissant notamment des corpus DIY monolingues :

> (2) J'utilise Sketch engine pour trouver des correspondances dans la langue cible afin de rendre mes traductions plus fluides.
>
> Utilisation de corpus à l'aide d'AntConc surtout pour me familiariser avec la terminologie d'un client/sujet particulier en compilant des articles ou fichiers pertinents.
>
> J'utilise principalement des corpus pour les très gros projets ou pour les projets très techniques.
>
> Je compile très souvent des mini corpus en langue cible pour des traductions techniques et je consulte très souvent des corpus en ligne en langues source et cible.
>
> J'utilise des corpus au quotidien, ceux disponibles en ligne.
>
> J'utilise principalement des corpus pour trouver des collocations ou me familiariser avec le style d'un domaine donné.

> Je trouve les corpus extrêmement utiles pour améliorer la fluidité de la langue cible.
>
> (3) Les délais demandés par les clients ne me permettent quasiment jamais de prendre le temps de consulter un corpus. En outre, l'utilisation de corpus de qualité en ligne est payante.
>
> En vrai, on n'a pas le temps de faire des concordanciers, parce qu'on n'a pas le temps de chercher des textes pour les deux langues. Je n'ai jamais utilisé cela depuis 2014, pourtant ça avait l'air top…
>
> Du fait des clients et des deadlines serrées, il m'arrive peu souvent d'avoir recours aux corpus à proprement parler (à l'exception des mémoires de traduction qui ont une fonction semblable), car le travail demandé repose plus sur la quantité que sur la quantité souvent.
>
> J'ai toujours eu tendance à considérer la compilation et l'exploitation de corpus comme des opérations fastidieuses et chronophages.
>
> Les corpus ne sont généralement pas utilisés en agence/entreprise, peut-être parce que les ressources fournies par les clients sont déjà suffisantes.
>
> Je n'ai pas le temps d'utiliser des corpus dans mon travail quotidien, et je n'en ai pas forcément l'utilité.

2.3. L'influence du statut et de la spécialisation ?

À la lumière des remarques libres formulées par les répondants, nous avons souhaité voir si le statut (exercice du métier sous statut indépendant vs en agence ou en entreprise) mais aussi l'existence d'une ou plusieurs spécialisation(s) avaient une influence sur l'utilisation actuelle des outils de corpus. Il semble bien que cela soit le cas, puisque chez les répondants travaillant en agence ou en entreprise, 43 % (3/7) utilisent au moins un type de corpus dans le cadre de leur activité (en ligne ou DIY), une proportion qui monte à plus de 66 % (23/35) chez ceux qui exercent en tant qu'indépendants. La liberté des outils utilisés du côté des indépendants n'est sans doute pas étrangère à ce résultat. De façon encore plus marquée, 71 % des répondants (20/28) ayant une spécialisation utilisent au moins 1 type de corpus, proportion qui n'est que de 43 % (6/14) chez ceux qui n'en ont pas. Si l'on se concentre sur les corpus monolingues DIY, on constate que c'est chez les indépendants qui ont une spécialisation qu'ils sont le plus utilisés, même si cela reste minoritaire : 7 indépendants déclarent avoir recours à de tels corpus contre 1 seul non indépendant ; 6 professionnels ayant une spécialisation exploitent des corpus DIY contre 2 n'en ayant pas. Même si nos résultats ne portent que sur un faible échantillon, il semble que l'utilisation des corpus soit particulièrement utile en cas de spécialisation (voir Eddy, 2020 pour un témoignage intéressant en traduction juridique), où les textes de travail relèvent de domaines et de genres bien précis (on parle alors de « minilectes »). À l'heure où l'on parle de plus en plus pour les professionnels d'« hyperspécialisation », ce résultat nous semble particulièrement intéressant.

2.4. Quid du marché en général ?

Les acteurs et actrices du marché des services linguistiques n'ayant pas nécessairement suivi une formation universitaire, il peut être intéressant de confronter les résultats de notre enquête de terrain à des résultats plus généraux. Ainsi, si l'on s'intéresse aux enquêtes annuelles ELIS (*European Language Industry Surveys*)[8] menées au niveau européen afin de connaître l'état du secteur des services linguistiques et leur perception par les différents acteurs, il est possible de connaître l'utilisation par les professionnels des différents outils disponibles aujourd'hui.

Ainsi les enquêtes 2022, 2023, et 2024 montrent que les outils de corpus sont bien utilisés, mais de façon toutefois limitée si l'on compare avec la fréquence d'utilisation des autres outils, ce qui rejoint nos propres constatations. L'enquête 2022, la plus précise, fait état d'une utilisation quotidienne et régulière chez 1 personne sur 10 parmi les indépendants interrogés (n=745), avec une utilisation occasionnelle chez 25 % de répondants supplémentaires, soit un total d'environ un tiers des répondants qui utilisent des outils de corpus (contre des proportions allant de 70 à 90 % pour une utilisation, quelle que soit la fréquence, d'outils comme la TAO, la TA, ou encore les extracteurs terminologiques). Les enquêtes 2023 et 2024, moins précises, indiquent une utilisation chez environ 20 % des indépendants interrogés (n=636 pour 2023, n=919 en 2024), contre plus de 70 % pour les mémoires de traduction, entre 45 et 55 % pour la traduction automatique. Aucune distinction n'est toutefois faite selon le type de corpus (en ligne/DIY), l'ensemble des outils étant nommés « *corpus analysis* ».

Des études universitaires ont également montré ce recours limité aux corpus par les professionnels de la traduction. Par exemple Gallego-Hernández (2015) a mis au jour, au sein d'une population de 526 professionnels installés en Espagne, une utilisation occasionnelle pour 30 % et fréquente pour 18 % d'entre eux. Si cette étude remonte maintenant à plusieurs années, les dernières enquêtes ELIS montrent que malgré une évolution, l'utilisation de données linguistiques réunies en corpus reste marginale par rapport à d'autres outils.

2.5. La formation continue

Au-delà de la formation initiale, c'est au cours de formations continues qu'il est possible d'acquérir des compétences en matière de compilation et d'exploitation de bases de données linguistiques. C'est ce que propose notamment la Société française des traducteurs (SFT), avec une formation intitulée « Les corpus pour la traduction et l'interprétation »[9] que nous dispensons depuis 2019 et au sein de laquelle nous formons à la compilation et l'exploitation des corpus DIY spécialisés. Nous avons souhaité interroger les professionnels ayant participé à ces formations d'une durée d'une journée (au nombre de 5 entre 2019 et 2023, à raison d'une dizaine de participants à chaque session), afin de dresser un bilan similaire à ce qui a été fait pour les anciens étudiants de la formation universitaire TSM. Un questionnaire à remplir en ligne leur a été transmis par voie

[8] Les enquêtes sont disponibles ici : https://elis-survey.org/

[9] Si nous ne parlons ici que de traduction, il est important de noter que les outils de corpus peuvent également être utiles au travail des interprètes.

électronique, et nous avons obtenu 26 réponses. Les répondants ont participé de façon plutôt égale aux différentes sessions, avec une moyenne de 5,2 répondants par session (valeur médiane = 4). Il s'agissait exclusivement de professionnels exerçant de façon indépendante, avec au moins une spécialisation dans la très grande majorité des cas (9 cas sur 10) : juridique, finance, marketing, informatique, médical, communication d'entreprise, tourisme, musique, aérospatial, nutrition, énergies renouvelables…

Cette enquête de terrain nous apprend qu'avant la formation, trois quarts des répondants n'avaient aucune compétence en compilation et exploitation de corpus, alors que plus de 80 % estiment que la découverte de ces outils a toute sa place dans une formation initiale. S'agissant des compétences acquises lors de la formation et conservées par la suite, on constate comme pour la formation initiale une érosion globale des compétences. À l'issue de la formation, les répondants estiment en effet avoir été capables d'utiliser les corpus en ligne pour améliorer la compréhension d'un texte source (34,6 %) ou améliorer la qualité de leur rédactionnel en langue cible (65,4 %) ; ils estiment par ailleurs avoir été en mesure de compiler et d'exploiter des corpus DIY à exploiter hors ligne à l'aide d'un concordancier pour effectuer les mêmes tâches à hauteur de 46,2 % et 69,2 % et pour compiler un glossaire à hauteur de 46,2 %. Ces résultats, schématisés par la Figure 4, sont prudents et contrastent fortement avec ceux des étudiants à l'issue de leur formation, où les compétences étaient considérées comme acquises de façon bien plus généralisée. Les compétences, n années après la formation (entre 1 et 4), sont par ailleurs en recul comme cela est le cas pour la formation initiale, à une exception près : les compétences relatives à l'utilisation de corpus spécialisés DIY hors ligne sont conservées pour 38,5 %, 50 % et 19,2 % des répondants, le plus fort recul concernant la compilation de glossaires. S'agissant des corpus en ligne, seul 1 répondant sur 2 estime être capable de les utiliser pour améliorer la fluidité en langue cible (contre 2 sur 3 à l'issue de la formation) ; en revanche, on note une augmentation lorsqu'il s'agit d'améliorer la compréhension de la langue source. On constate là encore que les compétences relatives aux corpus DIY monolingues sont difficiles à acquérir et à conserver.

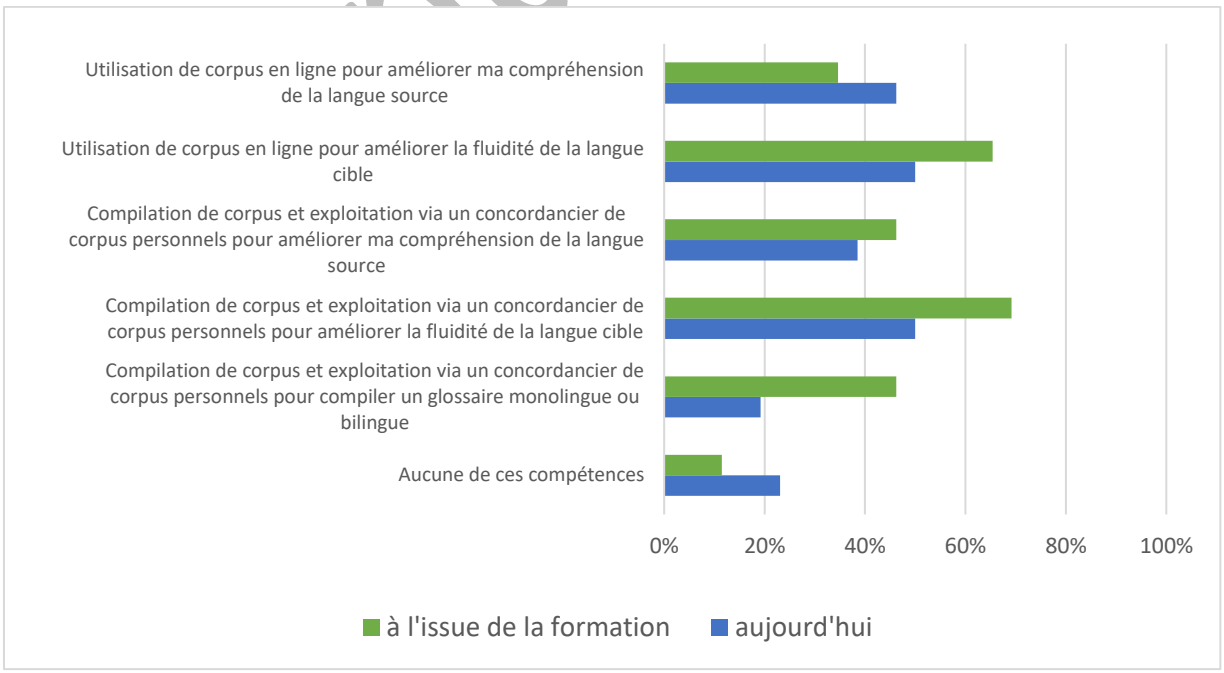


Figure 4. Compétences considérées comme acquises après une formation continue

S'agissant de la mise en œuvre des compétences et donc de l'utilisation réelle des outils, les corpus en ligne ont été utilisés depuis la formation par 35 % et 40% des répondants pour améliorer la compréhension de la langue source et la fluidité de la langue cible respectivement. Du côté des corpus DIY monolingues, ils ont été utilisés comme outils d'aide à la compréhension de la langue source (40 % des répondants), à la rédaction en langue cible (60 %), et pour compiler un glossaire (15 %). Par rapport à l'utilisation déclarée par les professionnels ayant été formés en formation initiale, on constate en moyenne une utilisation bien plus fréquente des outils de corpus après une formation continue (voir Figure 5), en particulier pour les corpus DIY spécialisés. Le nombre d'années suivant la formation peut naturellement avoir une influence (l'utilisation chez les stagiaires ayant suivi une formation continue ne durera peut-être pas dans le temps), mais ces résultats laissent penser que les compétences relatives à l'exploitation et en particulier à la compilation de corpus sont particulièrement pertinentes une fois l'activité professionnelle lancée, ce qui n'est pas sans rappeler les résultats concernant la formation initiale, qui montrait une utilisation plus fréquente lorsqu'existe une spécialisation. Il reste toutefois important de remarquer que ces résultats relatifs à l'utilisation contrastent avec l'acquisition et la conservation estimées des compétences.

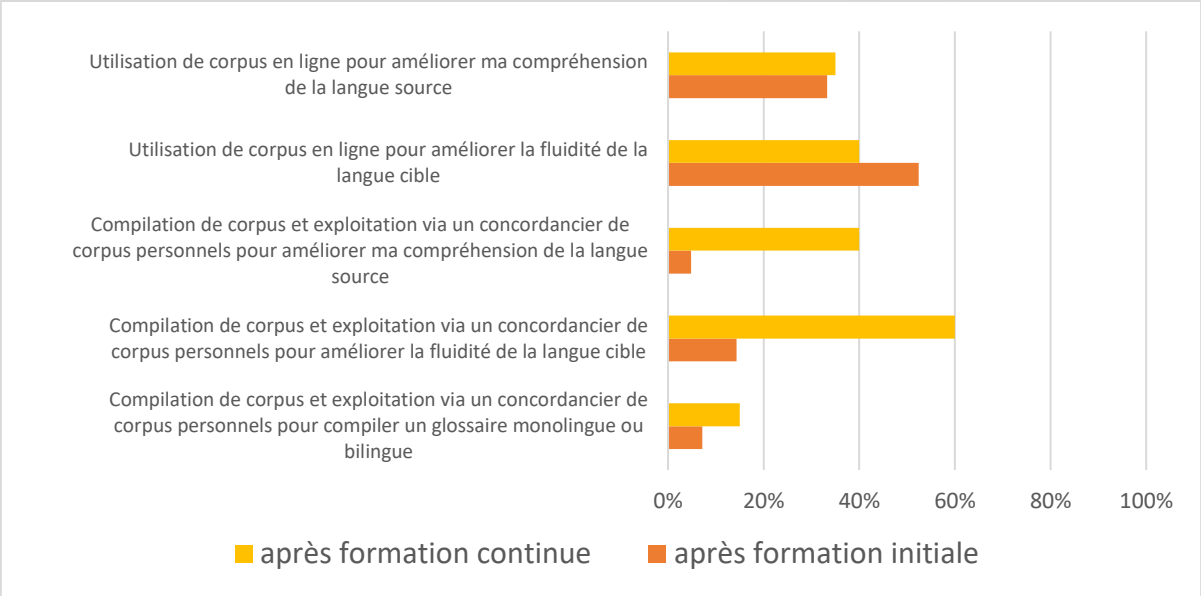


Figure 5. Comparatif de la fréquence d'utilisation (en %) des outils de corpus après la formation continue et la formation initiale

3. Nos propositions

Suite aux différents constats évoqués ci-dessus qui témoignent d'une utilisation qui reste faible malgré une acquisition de compétences plutôt réussie, nous souhaitons formuler ici quelques propositions visant à optimiser la formation aux outils de corpus pour un public d'étudiants en traduction spécialisée, et à combler le fossé souvent décrié qui peut exister entre formation et activité professionnelle. En particulier, nous proposons un repositionnement stratégique dans la façon dont les corpus électroniques, notamment spécialisés, sont présentés dans les formations universitaires aux métiers de la traduction.

3.1. Concentrer l'enseignement sur le développement d'une « *data literacy* » et une approche globale des outils

Dans la mesure où les corpus sont présents dans de nombreux outils utilisés en traduction, il peut être intéressant d'axer l'enseignement sur une éducation aux données et sur leur importance capitale dans les résultats affichés, tout en adoptant une approche globale des différents outils. Ainsi, des différences de résultats pour une même recherche selon l'outil utilisé (dictionnaires en ligne enrichis de type Linguee, traducteur en ligne de type DeepL, mais aussi l'internet en général fréquemment utilisé pour des recherches linguistiques malgré les problèmes de qualité que nous avons évoqués) peuvent mettre au jour l'influence qu'ont les données sous-jacentes sur les informations obtenues, en particulier s'agissant de terminologie spécialisée. Les limites de certaines ressources peuvent alors être identifiées, déclenchant le besoin pour des données linguistiques plus contrôlées, compilées par ses propres soins selon des critères précis et ne contenant que des textes en adéquation avec le projet de traduction en cours, en d'autres termes des corpus DIY (cf. notre approche pour la traduction du document sur les plugins ChatGPT, voir section 1.3). Il devient alors indispensable d'aborder toute une série d'enjeux techniques, comme l'adéquation, la fiabilité, la représentativité, et la pertinence des données, mais aussi l'existence de biais (algorithmiques, discriminatoires) dans les résultats présentés, soit toute une série d'enjeux communs à l'ensemble des outils faisant intervenir des données.

Ce type d'éducation aux données a été proposée dans le cadre du développement d'une utilisation professionnelle des outils de traduction automatique. Partant du constat de leur omniprésence et du fait que la traduction professionnelle connaît une numérisation et une « datafication » de grande ampleur, Krüger & Hackenbuchner (2022) proposent ainsi une approche didactique qui mette en avant ce rôle crucial des données exploitées par les outils de traduction automatique, avec pour objectif le développement d'une « *data literacy* », définie comme une « capacité à collecter, gérer, évaluer, et exploiter des données de façon critique » (Ridsdale *et al.*, 2015 : 11, cité par Krüger & Hackenbuchner, 2022 : 392, nous traduisons). Le développement d'une telle compétence passe alors par une connaissance de l'environnement au sein duquel évoluent les données (*data context*), une capacité à développer une stratégie pour les identifier et les gérer (*data planning*), des compétences techniques relatives à leur production (*data collection/production*) et à leur évaluation (*data evaluation*) ainsi qu'à leur utilisation dans le cadre de la prise de décision (*data use*) (voir Krüger & Hackenbuchner 2022 : 396-402 pour une présentation plus complète). Cette approche nous paraît particulièrement pertinente et permet un repositionnement stratégique de la compilation et de l'exploitation des outils de corpus : plutôt que de présenter de nouveaux outils de façon cloisonnée, il s'agit alors d'uniformiser l'approche vis-à-vis de la gamme d'outils à disposition des traducteurs (des dictionnaires en ligne aux outils génératifs qui s'appuient sur l'intelligence artificielle) en les présentant comme des bases de données interrogeables via différents types d'interfaces (site internet, logiciel dédié) dans le but d'obtenir différents types d'informations.

3.2. Remplacer le terme « corpus » par « base de données (linguistiques) »

Dans la continuité de l'approche axée sur les données elles-mêmes, nous proposons, de façon potentiellement controversée nous en convenons, de remplacer le terme « corpus » par « base de données (linguistiques) ». Le mot « corpus » semble en effet poser un certain nombre de problèmes car perçu (à tort) comme étant réservé à la recherche universitaire en linguistique ou en traductologie, loin de toute préoccupation professionnelle. On constate en effet que ce terme est souvent absent du milieu professionnel : nous avions ainsi souligné son absence dans les offres d'emploi et les rapports de stage des étudiants (Loock, 2023). Ceci semble paradoxal à une époque où la question des données est abondamment discutée en lien avec les nouveaux outils d'IA générative. On constate d'ailleurs une utilisation du terme « donnée » bien plus répandue. Les outils de traduction assistée par ordinateur (TAO) proposent par ailleurs des fonctionnalités semblables à ce qui peut être fait avec des corpus (outils *Concordance Search* chez RWS Trados Studio ou *LiveDocs* chez memoQ) mais sans qu'apparaisse systématiquement le mot « corpus ». Il semble donc que pour combler l'écart terminologique avec le monde professionnel, le terme « base de données (linguistiques) » puisse être préféré au terme « corpus ».

Ceci n'est pas sans poser problème et présente des inconvénients : au-delà du risque de rupture avec les concepts de la linguistique de corpus qui sont incontournables (représentativité, concordance, biais), certains outils en ligne ou logiciels utilisent le terme « corpus ». Nous pensons par exemple aux sites et outils mentionnés dans la section 1 (site english-corpora, portail Sketch Engine, concordancier AntConC). La faisabilité et la pertinence de ce glissement terminologique restent donc à démontrer, mais nous avons pu faire l'expérience que le terme « base de données (linguistiques) » attire davantage l'attention de professionnels.

3.3. Distinguer clairement outils d'aide à la traduction et outils de recherche

Toujours dans la continuité des deux propositions précédentes, il nous semble indispensable d'opérer une distinction très claire dans les enseignements entre l'utilisation des corpus comme outils d'aide à la traduction et comme outils d'analyse à des fins de recherche, ce qui n'est pas toujours le cas. Il existe en effet un biais dans la recherche en traductologie de corpus, où les textes rassemblés en corpus le sont bien souvent à des fins de recherche (Mikhailov, 2022). Nous avons pu nous-mêmes le constater grâce à une recherche effectuée via l'outil Google Scholar (Loock, 2023), dans les publications en langue anglaise entre 1990 et 2023 avec certains mots clefs (*translation, corpus/corpora, translation studies, aid(s), tool(s), translator(s)*). Celle-ci révèle que si les termes *translation* et *corpus* apparaissent dans les titres de 2160 publications, les termes *translator(s), aid(s)*, ou encore *tool(s)* n'apparaissent que de façon marginale (respectivement 27, 7, et 50 fois). D'après Mikhailov (2022), ce biais se retrouve dans la formation des futurs traducteurs : il montre en

effet dans le cadre de son étude sur 71 formations universitaires que la formation aux corpus électroniques se fait en premier lieu en lien avec la recherche en traductologie. De façon intéressante, un constat similaire a été fait pour l'enseignement des langues à partir de données ou DDL en anglais (pour *data-driven learning*) dans Crosthwaite & Baisa (2024), qui prône l'utilisation d'outils simples et ergonomiques si l'on souhaite qu'un tel type d'enseignement se démocratise, alors que les outils actuels permettant l'exploration de corpus sont parfois complexes, orientés vers la recherche, et donc intimidants pour enseignants et apprenants, qui ne parviennent pas à se les approprier.

3.3. Mettre en place une approche métacognitive

Enfin, en adéquation avec les propositions formulées ci-dessus, nous proposons de mettre en place systématiquement avec les étudiants une approche métacognitive, à savoir une réflexion menée par les étudiants eux-mêmes sur ce qu'ils apprennent d'un enseignement et sur ce qui pourra être réutilisé par la suite, en les invitant à réfléchir sur leurs savoirs et compétences en matière de bases de données linguistiques[10]. La métacognition, ou connaissances à propos de ses propres connaissances, consiste alors à amener les étudiants à verbaliser le lien entre ce qui leur est enseigné et leur future pratique professionnelle. C'est une approche que nous avons mise en place de façon expérimentale ces dernières années en seconde année de formation de master dans le cadre de l'enseignement consacré aux outils de corpus, en partant du constat que les étudiants semblaient éprouver des difficultés à faire ce lien. À 15 minutes de la fin de chaque séance, nous les invitons à rédiger une réponse courte (entre 5 et 10 lignes) à la question suivante : *Qu'avez-vous retenu du cours d'aujourd'hui qui soit applicable dans votre future vie professionnelle ?* Des expressions du type « nous avons appris/vu que », « j'ai découvert », « je sais » sont proscrites, au profit d'expressions du type « dans le cadre de ma pratique professionnelle, je serai en mesure de/je pourrai/je devrai… ». Voici en (4) 3 exemples de réponses obtenues à l'issue d'une des séances (aucune modification n'a été apportée, c'est en revanche nous qui soulignons certains éléments) :

> (4) Lors du cours d'aujourd'hui, j'ai retenu de nombreux aspects que je vais évoquer. D'abord, compiler une base de données linguistiques dans ma future vie professionnelle me permettra de **gagner du temps seulement si je sais me servir des outils en question**. Ensuite, j'ai conscience que si je compile souvent des bases de données linguistiques et que cela fait partie de mon quotidien dès aujourd'hui (mais aussi dès le M1), **j'ai la possibilité d'avoir énormément de ressources que je dois conserver** et sur lesquelles je pourrai revenir plus tard. Je sais toutefois qu'il convient d'**utiliser ces outils et ces logiciels de façon récurrente** pour pouvoir se les approprier et cela m'évitera de perdre du temps à l'avenir. En effet, dans ma future carrière de traductrice, à la fin de la journée de travail, ce qui importera ce sera le nombre de mots traduits.
>
> Ce qui m'a le plus marqué pendant ce cours, c'est le **rapport entre les bases de données linguistiques et la spécialisation**. En effet, pour moi, la spécialisation dans un certain domaine a toujours été une évidence puisque depuis le début du Master, on nous répète qu'il est important de se spécialiser. Pourtant, je n'ai jamais

[10] Voir Flavell (1979) pour une définition détaillée de l'approche métacognitive, pour qui il s'agit de comprendre et de contrôler ses propres processus cognitifs.

> trouvé un domaine de spécialisation qui pourrait réellement m'intéresser, ou dans lequel j'ai assez de connaissances. Mais, grâce aux corpus, j'ai compris que pour se lancer dans un domaine de spécialisation, il n'y a pas besoin de connaître tout sur le sujet dès le début, puisque **les outils qui sont mis à notre disposition peuvent nous permettre de compiler de nombreux textes spécialisés afin de ne pas passer des heures, voire des jours, à lire des dizaines de documents**, mais plutôt de faire des recherches précises afin de ne pas perdre de temps et d'avoir directement des informations qui nous sautent aux yeux, comme par exemple l'article qui va avec un terme précis. Je pense donc qu'il sera plus facile pour moi de me projeter dans une spécialisation grâce aux bases de données linguistiques.
>
> Je me suis rendue compte que certaines choses évoquées pendant le cours d'aujourd'hui pourraient me servir pour ma future vie professionnelle. Lorsque je traduirai, je devrai avoir conscience que **la langue source va forcément influencer mon texte cible**. Je devrai donc **mettre des stratégies en place**, comme par exemple prendre du recul ou utiliser des synonymes, entre autres. Aujourd'hui, j'ai également appris qu'en traduction ce n'est pas toujours une bonne chose d'expliciter puisqu'il faut toujours rester conforme au texte source et ne pas s'en éloigner. En outre, j'ai compris que dans ma future carrière de traductrice indépendante, **je ne devrai pas avoir peur d'utiliser des formes non standard**. Il convient de se démarquer afin d'**apporter sa touche personnelle** et cela évite ainsi de perdre en créativité. Pour résumer les points mentionnés, en tant que traducteur.trice, il est nécessaire de trouver un juste milieu : rester proche du texte cible sans toutefois traduire littéralement. Enfin, **il ne faut pas avoir peur du taux de foisonnement** qui est indispensable et qui n'est donc pas uniquement le résultat de l'explicitation.

Conclusion

Dans cet article, nous avons souhaité dresser un bilan critique de dix années d'un enseignement consacré au développement de compétences en matière de compilation et d'exploitation de corpus, que nous souhaitons renommer « bases de données linguistiques », au sein d'une formation de master en traduction spécialisée. Cette prise de recul, au moyen de différentes analyses et enquêtes de terrain, nous a permis de dresser un certain nombre de constats. La maîtrise de ces outils, et notamment les corpus dits « DIY » en langue cible originale avec pour but l'amélioration de la qualité rédactionnelle en langue cible, n'est pas chose facile et prend du temps, davantage que pour d'autres outils comme les logiciels de TAO que les étudiants s'approprient plus rapidement. Les compétences acquises s'érodent avec le temps une fois la formation, initiale ou continue, terminée, même si ce recul reste limité. Il n'en demeure pas moins que le recours aux bases de données linguistiques, et notamment aux corpus monolingues spécialisés, reste marginal dans la pratique professionnelle. Malgré tout, on constate l'utilité de ces outils puisqu'ils amènent les utilisateurs à réviser leurs traductions, en y apportant des améliorations que d'autres outils n'ont pas permises. À partir de ces constats, nous avons souhaité émettre un certain nombre de propositions, pour un repositionnement stratégique de l'enseignement qui s'articule autour des données linguistiques elles-mêmes, avec une approche globale de l'ensemble des outils qui exploitent des données linguistiques. D'autres propositions, comme l'abandon du terme « corpus » ou l'approche métacognitive, visent à réduire le fossé entre formation et pratique professionnelle.

Références :

ANTHONY Laurence (2023). AntConc (Version 4.2.4). Disponible sur https://www.laurenceanthony.net/software.

BEVENDORFF Janek, WIEGMANN Matti, POTTHAST Martin & STEIN Benno (2024), « Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines». N. Goharian *et al.* (dir.), *Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science*, Springer, 14610, <https://doi.org/10.1007/978-3-031-56063-7_4>.

BOWKER Lynne & PEARSON Jennifer (2002), *Working with Specialized Language: A Practical Guide to Using Corpora*. Londres : Routledge ELIS Research group.

European Language Industry Surveys 2022, 2023, 2024. Disponibles sur <https://elis-survey.org>.

CROSTHWAITE, Peter & BASA Vít (2024), « A user-friendly corpus tool for disciplinary data-driven learning: Introducing *CorpusMate* », *International Journal of Corpus Linguistics,* Online First*,* <https://doi.org/10.1075/ijcl.23056.cro>.

EDDY, Charles (2020), « Legal Translation and Corpora: A Crash Course in Monolingual DIY Corpora for Legal Translators», billet de blog publié sur le site C. Eddy Traductions, <https://ceddytraductions.fr/articles/2020-03-24_diy-corpora-in-legal-translation.shtml>

FLAVELL John H. (1979), « Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry », *American Psychologist*, *34*(10), 906-911, <https://doi.org/10.1037/0003-066X.34.10.906>.

FRANKENBERG-GARCIA Ana (2015), « Training translators to use corpora hands-on: challenges and reactions by a group of 13 students at a UK university », *Corpora*, *10*(2), 351-380, <https://doi.org/10.3366/cor.2015.0081>.

FRÉROT Cécile & KARAGOUCH Lionel (2016), « Outils d'aide à la traduction et formation de traducteurs : vers une adéquation des contenus pédagogiques avec la réalité technologique des traducteurs » *ILCEA, 27* (en ligne), <https://doi.org/10.4000/ilcea.3849>.

GALLEGO-HERNÁNDEZ Daniel (2015), « The use of corpora as translation resources: a study based on a survey of Spanish professional translators », *Perspectives: Studies in Translatology, 23*(3), 375-91, <https://doi.org/10.1080/0907676X.2014.964269>.

GIAMPIERI Patrizia (2021), « Can Corpus Consultation Compensate for the Lack of Knowledge in Legal Translation Training? », *Comparative Legilinguistics, 46*(1), 5-35, <http://dx.doi.org/10.2478/cl-2021-0006>.

KRÜGER Ralph & HACKENBUCHNER Janiça (2022), « Outline of a didactic framework for combined data literacy and machine translation literacy teaching », *Current Trends in Translation Teaching and Learning E, 9*, 375-432. <https://doi.org/10.51287/cttl202211>.

KÜBLER Natalie (2003), « Corpora and LSP translation», F. Zanettin, S. Bernardini, & D. Stewart (dir.), *Corpora in Translator Education*, St Jerome, 25-42.

KÜBLER Natalie, MESTIVIER Alexandra & PECMAN Mojca (2022), « Using Comparable Corpora for Translating and Post-editing Complex Noun Phrases in Specialized Texts: Insights from English to French Specialised Translation », S. Granger & M.-A. Lefer (dir.), *Extending the Scope of Corpus-Based translation Studies*, Bloomsbury Publishing, 237-266.

LOOCK Rudy (2016a), « L'utilisation des corpus électroniques chez le traducteur professionnel : quand ? comment ? pour quoi faire ? », *ILCEA* 27 (en ligne), <https://doi.org/10.4000/ilcea.3835>.

LOOCK Rudy (2016b), *La Traductologie de corpus*, Villeneuve d'Ascq : Presses Universitaires du Septentrion.

LOOCK Rudy (2023, janvier), « From "stuff for linguists" to professional translation tools: the complicated relationship between translators and corpora », communication présentée au colloque international *Convergence: human-machine integration in translation and interpreting*, University of Surrey, Royaume-Uni.

MIKHAILOV Mikhail (2022), « Text corpora, professional translators and translator training », *The Interpreter and Translator Trainer, 16*(2), 224-246, <https://doi.org/10.1080/1750399X.2021.2001955>.

Référentiel de compétences 2022 du réseau *European Master's in Translation* (EMT) de la Commission européenne. Disponible sur <https://commission.europa.eu/system/files/2023-01/emt_competence_fwk_2022_fr.pdf>.

RIDSDALE Chantel, ROTHWELL James, SMIT Michael, ALI-HASSAN Hossam, BLIEMEL Michael, IRVINE Dean, KELLEY Daniel, MATWIN Stan & WUETHERICK Bradley (2015), *Strategies and BestPpractices for Data Literacy Education. Knowledge synthesis report*. Dalhousie University, Canada.

ROTHWELL Andrew, MOORKENS Joss, FERNÁNDEZ-PARRA María, DRUGAN Joanna & AUSTERMUEHL Frank (2023), *Translation Tools and Technologies*. (1e éd.), Londres : Routledge. <https://doi.org/10.4324/9781003160793>.

SÁNCHEZ-GIJÓN Pilar (2009). « DIY corpora in the Specialised Translation Course», A. Beeby, P. Rodríguez-Inés, & P. Sánchez-Gijón (dir.), *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Amsterdam/Philadelphie : John Benjamins, 109-128.

SCOTT, Juliette (2012). « Can genre-specific DIY corpora, compiled by legal translators themselves, assist them in learning the linguo of legal subgenres? », *Comparative Legilinguistics, 12*, 87-100.

THOMPSON Brian, DHALIWAL Mehak Preet, FRISCH Peter, DOMHAN Tobias & FEDERICO Marcello. A paraître. « A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism». Disponible sur <https://arxiv.org/abs/2401.05749>.

ZANETTIN Federico, BERNARDINI Silvia & STEWART Dominic (2003), *Corpora in Translator Education*, Londres : Routledge.

Annexe : Enquête sur votre utilisation actuelle des outils de corpus

1. J'appartiens à la promotion diplômée en :

2014
2015
2016
2017
2018
2019
2020
2021
2022

2. Mes paires de langues sont

anglais-français
allemand-français
espagnol-français
italien-français
chinois-français
suédois-français
russe-français
néerlandais-français
Autre :

3. J'exerce aujourd'hui :

en agence de traduction
en tant qu'indépendant(e)
Autre :

4. J'ai une ou plusieurs spécialisations

Oui
Non

5. Si vous avez répondu oui à la question précédente, pouvez-vous préciser :

6. J'ai suivi l'enseignement consacré aux corpus en :

M1 et M2
M1 uniquement
M2 uniquement

7. Je dirais qu'à l'issue de cet enseignement, j'avais acquis les compétences suivantes (plusieurs réponses possibles) :

Utilisation de corpus en ligne pour améliorer ma compréhension de la langue source
Utilisation de corpus en ligne pour améliorer la fluidité de la langue cible
Compilation de corpus et exploitation via un concordancier (AntConc) de corpus personnels pour améliorer ma compréhension de la langue source
Compilation de corpus et exploitation via un concordancier (AntConc) de corpus personnels pour améliorer la fluidité de la langue cible
Compilation de corpus et exploitation via un concordancier (AntConc) de corpus personnels pour compiler un glossaire monolingue ou bilingue
Aucune de ces compétences

8. Je dirais qu'aujourd'hui j'ai les compétences suivantes :

Utilisation de corpus en ligne pour améliorer ma compréhension de la langue source
Utilisation de corpus en ligne pour améliorer la fluidité de la langue cible
Compilation de corpus et exploitation via un concordancier (AntConc) de corpus personnels pour améliorer ma compréhension de la langue source
Compilation de corpus et exploitation via un concordancier (AntConc) de corpus personnels pour améliorer la fluidité de la langue cible
Compilation de corpus et exploitation via un concordancier (AntConc) de corpus personnels pour compiler un glossaire monolingue ou bilingue
Aucune de ces compétences

9. Dans ma vie professionnelle, il m'arrive d'accomplir les tâches suivantes :

Utilisation de corpus en ligne pour améliorer ma compréhension de la langue source
Utilisation de corpus en ligne pour améliorer la fluidité de la langue cible
Compilation de corpus et exploitation via un concordancier (AntConc) de corpus personnels pour améliorer ma compréhension de la langue source
Compilation de corpus et exploitation via un concordancier (AntConc) de corpus personnels pour améliorer la fluidité de la langue cible
Compilation de corpus et exploitation via un concordancier (AntConc) de corpus personnels pour compiler un glossaire monolingue ou bilingue
Aucune de ces tâches

10. Expression libre : quelle que soit votre réponse ci-dessus, pouvez-vous apporter des précisions sur votre (non) utilisation des corpus ?

11. Si vous utilisez des corpus en ligne, quels sites utilisez-vous ?

12. Je suis convaincu(e) que les corpus peuvent servir d'outils d'aide à la traduction

Oui
Non
Je ne sais pas

13. Question bonus : j'utilise des mémoires de traduction au sein d'un ou plusieurs logiciels de TAO :

Oui, souvent ou régulièrement
Oui, parfois
Oui, rarement
Non, jamais

14. Commentaires libres (optionnel)