ILCEA
Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie

27 | 2016
Approches ergonomiques des pratiques professionnelles et des formations des traducteurs

L'utilisation des corpus électroniques chez le traducteur professionnel : quand ? comment ? pour quoi faire ?

The Use of Electronic Corpora by Professional Translators: When? How? What for?

Rudy Loock

Édition électronique
URL : http://journals.openedition.org/ilcea/3835
ISSN : 2101-0609

Éditeur
UGA Éditions/Université Grenoble Alpes

Édition imprimée
ISBN : 978-2-84310-336-0
ISSN : 1639-6073

Référence électronique
Rudy Loock, « L'utilisation des corpus électroniques chez le traducteur professionnel : quand ? comment ? pour quoi faire ? », *ILCEA* [En ligne], 27 | 2016, mis en ligne le 08 novembre 2016, consulté le 09 janvier 2018. URL : http://journals.openedition.org/ilcea/3835

L'utilisation des corpus électroniques chez le traducteur professionnel : quand ? comment ? pour quoi faire ?

The Use of Electronic Corpora by Professional Translators: When? How? What for?

Rudy Loock

Je tiens à remercier dans un premier temps le public présent lors de la conférence « Traducteurs à l'œuvre : approches ergonomiques des pratiques professionnelles et des formations de traducteurs », dont les questions et remarques m'ont permis d'affiner ma réflexion. Je souhaite également remercier les étudiants des différentes promotions du Master « Traduction Spécialisée Multilingue » (TSM) de l'Université de Lille : les différentes expériences menées auprès d'eux m'ont permis de réfléchir aux problèmes ergonomiques que posent les corpus pour les traducteurs professionnels et d'adapter certaines méthodes de compilation et d'exploitation afin de mieux répondre à leurs besoins. En particulier, je remercie la promotion des étudiants de Master 2e année 2013-2014, auprès de qui le travail sur les nécrologies a été mené. Enfin, je remercie Laurence Anthony de m'avoir autorisé à utiliser des captures d'écran pour illustrer le fonctionnement du concordancier AntConc, ainsi que les deux relecteurs anonymes de cet article pour leurs judicieux conseils.

Introduction

1 Dans cet article, nous souhaitons nous pencher sur l'utilisation qui peut être faite des corpus électroniques, à savoir les banques de données linguistiques exploitables de façon automatique, par les traducteurs professionnels. Si cette question est bien loin d'être nouvelle (voir en France les travaux de Natalie Kübler notamment sur le sujet depuis la fin des années 1990), le recours aux corpus électroniques chez les traducteurs demeure

rare tandis que d'autres outils ont leur préférence, et les corpus n'ont pas acquis le statut d'outil d'aide à la traduction que nous souhaitons ici leur donner, pour des raisons que nous évoquerons dans la première partie de cet article. De façon paradoxale, les corpus sont pourtant omniprésents dans les outils utilisés par les traducteurs (p. ex. mémoires de traduction, outils de traduction automatique, ou encore sites internet très utilisés comme <www.linguee.com>), même si le terme « corpus » n'apparaît pas. Les traducteurs utilisent donc déjà certains types de corpus électroniques, mais sans toujours en avoir conscience. Parallèlement, d'autres types de corpus ne sont ni connus ni utilisés, comme les mini-corpus spécialisés en langue cible ou encore les corpus comparables de langues originales. L'une des raisons de cette non-utilisation (et parfois du rejet) des corpus comme outil d'aide à la prise de décision en traduction est clairement leur manque d'ergonomie : il existe de nombreux types différents de corpus électroniques, exploitables de façon parfois fastidieuse du fait de contenus et d'interfaces très hétérogènes tandis que d'autres outils semblent bien plus simples d'utilisation ; les corpus ne s'intègrent pas toujours très bien dans les outils déjà utilisés ; leur compilation et exploitation paraissent techniques, et donc réservées aux linguistes professionnels dans le cadre de travaux de recherche.

2 Ce que nous souhaitons montrer dans cet article, c'est que la compilation et l'exploitation de corpus ne sont ni fastidieuses ni chronophages et qu'elles peuvent permettre dans certains cas des gains de productivité et dans tous les cas de qualité. Nous souhaitons montrer que les corpus fournissent des informations que d'autres outils ne fournissent pas, notamment concernant l'usage de la langue, dont la prise en compte permet d'améliorer la fluidité des traductions. Enfin, l'objectif majeur de cet article est de formuler des propositions concrètes afin de résoudre les problèmes ergonomiques que pose l'utilisation des corpus électroniques : nous montrerons ainsi comment compiler de façon rapide un corpus spécialisé en langue cible et comment les différents corpus peuvent être utiles en fonction de l'étape du processus de traduction (de la compréhension du texte source à la phase de relecture). C'est en ce sens que nous souhaitons les considérer comme de véritables outils de TAO, bien que cette appellation soit traditionnellement réservée aux logiciels à mémoires de traduction.

3 L'article est organisé de la façon suivante. Nous évoquerons dans une première partie le paradoxe des corpus : bien qu'omniprésents, ils sont souvent invisibles et peu de traducteurs ont conscience d'en utiliser ; de la même manière, bien qu'ils puissent assister le traducteur dans sa tâche en augmentant productivité et qualité, celui-ci ne compile pas et n'exploite pas ses propres corpus. Nous expliquerons ce paradoxe par les problèmes d'ergonomie que pose l'utilisation des corpus électroniques. Dans une deuxième partie, nous expliquerons comment les corpus peuvent fournir des informations utiles au traducteur pour les différentes phases du processus de traduction : observation de la langue source, observation de la langue cible, recherche d'inspiration, évaluation de la qualité. En particulier nous nous attarderons sur deux exemples précis, issus d'expériences menées auprès d'étudiants en formation à l'Université de Lille : (i) la compilation d'un corpus « maison » spécialisé en langue cible et (ii) la compilation d'un corpus comparable de textes spécialisés. Nous montrerons dans les deux cas que l'exploitation de ces corpus fournit des réponses que d'autres outils (dictionnaires, glossaires en ligne...) ne fournissent pas, et tenterons de balayer au passage certaines idées reçues comme le caractère technique et chronophage de la compilation de corpus.

Les langues de travail seront l'anglais comme langue source et le français comme langue cible.

Le paradoxe des corpus

Omniprésence et invisibilité

4 Le premier paradoxe concernant les corpus est qu'ils sont à la fois omniprésents et invisibles. Ils se « cachent » en effet derrière les désormais omniprésentes mémoires de traduction utilisées au sein des logiciels de TAO. Les mémoires de traduction ne sont en fait rien d'autre que des corpus parallèles, mais elles ne sont jamais dénommées ainsi. Le marché, qui a imposé l'utilisation de ces mémoires, semble lui avoir préféré un terme non technique et plus explicite, « mémoires » faisant référence à ce qui a déjà été fait et peut être recyclé pour un nouveau projet de traduction. De la même manière, ce sont bien des banques de données linguistiques, et donc des corpus, que l'on trouve dans les outils de traduction automatique, qu'il s'agisse de traduction automatique statistique ou basée sur l'exemple. Ces corpus peuvent être des corpus parallèles ou des corpus monolingues de langue originale, comme c'est le cas pour l'outil SYSTRAN par exemple. Enfin, ce sont également des corpus que l'on trouve sur des sites internet très en vogue comme < www.linguee.com> ou <www.reverso.net> (fonctionnalité « Context »), très utilisés par les étudiants en formation et qui se présentent comme des dictionnaires nouvelle génération. Ainsi, Linguee propose en effet à la fois des correspondances lexicales comme un dictionnaire en ligne traditionnel et une série de phrases dans la langue de départ où apparaît le terme recherché tandis que la traduction de ces phrases apparaît sur la droite de l'écran, l'équivalent du terme recherché étant surligné en jaune pour une meilleure ergonomie. Même s'il est parfois compliqué de distinguer la phrase source de la phrase cible, il ne s'agit ni plus ni moins que d'un corpus parallèle pouvant servir de source d'inspiration à l'utilisateur. Nous pourrons également faire remarquer que la révolution du *big data* commence à populariser le recours aux banques de données linguistiques, dans le domaine de la culturomique (*culturomics*) ou de la linguistique légiste (*forensic linguistics*), voire dans le monde du journalisme.

5 En dépit de cette omniprésence, lorsque l'on interroge des traducteurs professionnels, ceux-ci déclarent ne pas utiliser de corpus dans leur travail, ce qui est faux comme nous venons de le voir : ils en utilisent bien mais ces derniers ne sont pas perçus de cette façon. Ceci ne serait aucunement problématique s'il ne s'agissait que d'un problème de dénomination. Mais l'on s'aperçoit aussi, notamment chez les étudiants en formation, que la non-perception de leur présence empêche de bien percevoir les possibilités et surtout les limites de ces outils : peu s'interrogent en effet sur la qualité des mémoires de traduction, sur leur provenance, sur la possibilité de déterminer le statut original/traduit du segment. De la même manière, peu s'interrogent sur les corpus utilisés dans le cadre des outils de traduction automatique : pourtant, la qualité de la traduction générée ne dépend pas que de l'algorithme utilisé mais aussi des banques de données auxquelles celui-ci est appliqué.

6 Également, une grande enquête menée en 2006 dans le cadre du projet MeLLANGE (Multilingual eLearning in LANGuage Engineering) a montré que les traducteurs professionnels interrogés (plus de 700 dans toute l'Europe) collectent bien ce que l'on peut appeler des corpus puisqu'ils rassemblent des textes spécialisés issus de l'internet

pour près de la moitié d'entre eux, mais la plupart les exploite soit en les lisant, soit en utilisant la fonction « Rechercher » de leur navigateur. Il s'agit pourtant là de compilation de corpus, mais sans exploitation automatisée.

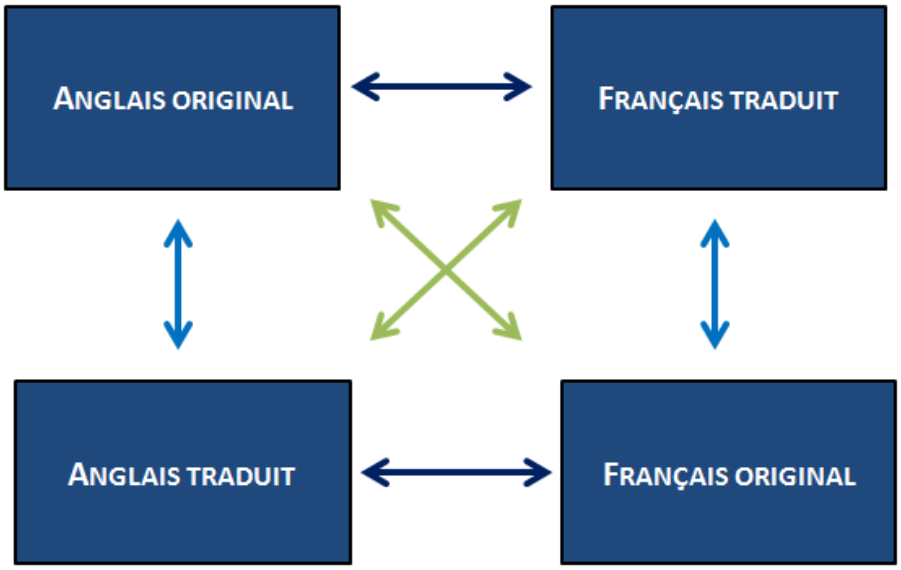
Explications

7 Comment expliquer ce paradoxe ? Tout d'abord, l'utilisation des corpus électroniques en traduction n'est pas enseignée dans tous les programmes de formation, même si les choses changent peu à peu, de plus en plus de formations accordant une place à l'enseignement de la compilation et de l'exploitation de corpus électroniques (voir p. ex. Beeby *et al.*, 2009 ; Bowker & Pearson, 2002 ; Frérot, 2010 ; Kübler, 2008, 2011 ; Kübler & Aston, 2010 ; Kunz *et al.*, 2010 ; Loock, 2014 ; Loock & Lefebvre-Scodeller, 2014 ; Zanettin *et al.*, 2003 ; voir aussi l'article de Cécile Frérot & Lionel Karagouch dans ce volume) et de nombreux ouvrages orientés vers la traduction ayant fait leur apparition ces dix dernières années (p. ex., Beeby *et al.*, 2009 ; Kübler, 2011 ; Mahadi *et al.*, 2010 ; Olohan, 2004 ; Zanettin, 2012). De ce fait, peu de traducteurs professionnels ont été sensibilisés à la question, ce qui fait que les corpus ne sont pas connus/utilisés sur le marché, et donc que les formations universitaires, toujours engagées à former leurs étudiants au plus près de la réalité du marché et toujours contraintes par les choix budgétaires, n'ont pas toutes la possibilité d'intégrer une formation à ces outils dans leur programme.

8 Mais surtout, et il s'agit là de l'objet de cet article, les corpus posent un véritable problème d'ergonomie. Une véritable confusion règne du fait de l'existence de nombreux types de corpus différents, utilisables seuls ou en association avec un/des autre(s) corpus. Ainsi, il existe des corpus monolingues de référence (p. ex. *British National Corpus*, *Corpus of Contemporary American English*, *Frantext*) ; des corpus multilingues (Europarl, *Wikipedia Corpus*) ; des corpus monolingues de langue traduite (p. ex. *Translational English Corpus*). Ces corpus peuvent alors être associés et former des corpus comparables ou parallèles selon que les sous-corpus considérés sont indépendants (p. ex. *British National Corpus vs Frantext* ; *British National Corpus vs Translational English Corpus*) ou que l'un des sous-corpus contienne des textes originaux et l'/les autre(s) sous-corpus la/les traduction(s) de celui-ci (p. ex. Europarl, mémoires de traduction), certains corpus pouvant contenir les deux (p. ex. *Oslo Multilingual Corpus*). Il s'agit alors d'effectuer des comparaisons inter-langagières (entre deux langues) ou intra-langagières (entre langue originale et langue traduite pour la même langue dans le cadre de la traductologie de corpus, voir Loock 2012a, 2012b), comme le schématise la figure 1, librement adaptée du célèbre schéma de Johansson et Oksefjell (1998 : 8) pour le *English-Norwegian Parallel* Corpus (ENPC). Certains corpus sont synchroniques tandis que d'autres comme le *Corpus of Historical American English* (COHA) sont diachroniques (Davies, 2010-) ; il est également possible de distinguer les corpus dits de référence et les corpus d'apprenants, qui contiennent des données produites par des non natifs de la langue ou, dans le cas de la traduction, des textes traduits par des apprentis traducteurs. Enfin, le traducteur a la possibilité de compiler ses propres corpus lorsque les corpus existants ne suffisent pas à ses besoins : on parle alors de corpus « maison », de corpus DIY (pour *do-it-yourself*), de corpus jetables (*disposable corpora*), de corpus *ad hoc*, de corpus virtuels, ou encore de mini-corpus spécialisés (Ahmad *et al.*, 1994 ; Mahadi *et al.*, 2010 : 15 ; Varantola, 2003 ; Zanettin *et al.*, 2003 : 7). Nous parlerons dans cet article de « corpus maison » ou encore de « corpus DIY » ; nous rejetons en revanche les appellations « jetable » ou encore « virtuel », car ces corpus ainsi compilés

n'ont absolument rien de virtuel et doivent être conservés, stockés pour de futures utilisations.

Figure 1. – Les différents types de corpus et de comparaisons possibles (anglais-français).



9 Ceci fait beaucoup de types de corpus différents, qui ne sont pas accessibles de la même façon : si certains le sont par le biais d'interfaces en ligne comme les corpus disponibles sur le site de la Brigham Young University (corpus.byu.edu/bnc/), d'autres doivent être exploités par le biais de logiciels spécifiques, à savoir des concordanciers comme AntConc par exemple (voir *infra*). L'accès aux corpus se fait donc de façon hétérogène, de façon parfois très peu ergonomique (voir le système de codes complexes à utiliser pour la partie catégorisée de Frantext), tandis que les contenus eux-mêmes ne sont pas comparables (ainsi Frantext contient principalement des textes littéraires et philosophiques tandis que les corpus monolingues de référence pour l'anglais contiennent davantage de registres comme les textes de presse ou les retranscriptions d'anglais oral). Il n'est alors pas étonnant que le recours aux corpus électroniques puisse paraître techniquement complexe et fastidieux, et par conséquent réservé aux analyses d'ordre linguistique et non à la traduction. En ce sens, comme le souligne Kübler (2011 : 63), l'utilisation de mémoires de traduction (qui sont pourtant des corpus, rappelons-le !) et des moteurs de recherche comme outils de vérification linguistique semble bien plus aisée. Parallèlement, comme le résume Frérot (2010), « l'impopularité des corpus chez les professionnels de la traduction tient au temps requis pour leur constitution et leur exploitation, et au fait que le gain de productivité n'est pas immédiat », problème auquel nous souhaitons apporter des solutions dans le cadre de cet article.

10 Le rôle du formateur est alors selon nous de SIMPLIFIER les choses : le traducteur professionnel, s'il est certes linguiste, n'a pas les mêmes objectifs que le linguiste professionnel qui effectue des recherches sur la langue. Ses besoins sont différents : aide à la prise de décision immédiate, contrôle de la qualité. Nous souhaitons donc dans cet article montrer comment cette simplification peut s'opérer, en insistant notamment sur le fait qu'il n'est pas nécessaire d'avoir suivi une formation poussée en linguistique, et en

adoptant une approche a-théorique. De la même manière, un protocole de compilation et d'exploitation de corpus spécialisés doit pouvoir se faire sans que l'utilisateur possède des compétences poussées en informatique. Enfin, nous souhaitons montrer que la compilation d'un corpus électronique à usage du traducteur ne devant pas répondre aux mêmes besoins que ceux du linguiste professionnel, celle-ci peut se faire de façon semi-automatisée sans que le corpus ait besoin d'être « nettoyé » ni étiqueté grammaticalement : nous développerons pour cela le concept de « corpus sale », en espérant ne pas choquer les linguistes qui passeraient par ici. Nous insisterons également sur le fait que la compilation et l'exploitation de corpus électroniques doivent être gratuites (ou tout au moins peu coûteuses) si l'on souhaite que les traducteurs se saisissent de ces outils.

11 Tout en faisant cela, nous souhaitons rationaliser l'utilisation des corpus électroniques en les associant aux différentes phases du processus de traduction, de la lecture du texte source à la phase de relecture, afin de montrer qu'ils peuvent être considérés comme des outils de TAO à part entière. Nous commencerons par évoquer brièvement la façon dont les corpus monolingues de référence peuvent être utilisés lors de la phase de compréhension du texte source ainsi que la façon dont les corpus parallèles peuvent servir de sources d'inspiration pour le traducteur. Nous nous attarderons en revanche sur deux utilisations beaucoup moins répandues : (i) l'observation de la langue cible à travers la compilation et l'exploitation de corpus spécialisés maison (corpus DIY) en langue cible (français) et (ii) la mise au jour de normes pour les « minilectes » à travers la compilation et l'exploitation d'un corpus comparable anglais orignal-français original.

La compréhension du texte source

12 Les corpus électroniques de référence s'avèrent particulièrement utiles pour la phase de compréhension du texte source, qui est la plupart du temps en langue étrangère. En complément des dictionnaires traditionnels, des corpus de référence comme le *British National Corpus* (BNC), qui contient 100 millions de mots d'anglais britannique contemporain (Davies, 2004-) ou le *Corpus Of Contemporary American English* (COCA), qui contient 520 millions de mots d'anglais américain contemporain (Davies, 2010-), apportent des informations précises concernant (i) le co-texte d'apparition d'un terme grâce aux très nombreux exemples en contexte, (ii) les phénomènes de collocation, de colligation et de prosodie sémantique, qui représentent des enjeux importants pour le traducteur (voir p. ex. Baker, 1992 ; Bowker, 2006 ; Kübler & Volanschi, 2012 ; Zanettin, 2012), (iii) les phénomènes de variation géographique ou liée au registre, soit toute une série de phénomènes relatifs à l'usage de la langue que les dictionnaires traditionnels ne peuvent pas couvrir (voir Loock, à paraître, pour toute une série d'exemples). Cette utilisation des corpus a fait l'objet de nombreuses expérimentations dans le cadre de l'apprentissage d'une langue seconde (voir à ce sujet les travaux d'Alex Boulton, p. ex. Boulton, 2008, 2012).

La recherche d'inspiration

13 Grâce aux corpus parallèles (collections de textes en langue A accompagnés de leur(s) traduction(s) en langue B), qui peuvent être bilingues (textes originaux en anglais ou en français, accompagnés de leur traduction dans l'autre langue) ou bien multilingues (par

exemple, textes originaux dans trois langues accompagnés de leurs traductions dans les deux autres langues), unidirectionnels (par exemple, le corpus contient des textes originaux en anglais et leur traduction en français) ou bi/multidirectionnels (le corpus contient des textes originaux dans plusieurs langues et leurs traductions dans les autres langues), le traducteur peut avoir accès à des traductions déjà réalisées dans le passé et s'en inspirer pour résoudre des problèmes de traduction dans le cadre de son propre projet. L'utilisation de ce type de corpus comme outil d'aide à la traduction est très répandue (et donc consensuelle) ; on les trouve sous la forme de mémoires de traduction, de corpus en ligne disponibles sur des sites internet comme les très en vogue www.linguee.com ou www.reverso.net, ou encore sous la forme de corpus DIY/maison exploitables par un concordancier (voir Barlow & Bowker, 2008 pour une comparaison entre mémoires de traduction et corpus parallèles DIY). En particulier, l'utilisation de ce que l'on appelle les mémoires de traduction dans le cadre des logiciels de TAO, qui ne sont rien d'autre que des corpus parallèles, est très répandue aujourd'hui chez les traducteurs professionnels.

L'observation de la langue cible

Objectifs : invisibilité

14 L'une des difficultés auxquelles le traducteur professionnel est confronté est qu'au-delà du sens du texte de départ à restituer (fidélité), il doit également produire un texte en langue cible qui soit naturel (fluidité). Si l'on exclut les cas de traduction littéraire ou les cas de traduction tendant vers l'« étrangéisation » (*foreignization* en anglais, voir Venuti, 1995), le traducteur cherchera donc à produire une traduction dans une langue qui soit la plus naturelle possible, afin de faire passer le texte traduit pour un texte écrit en langue originale. Le traducteur recherche alors l'invisibilité, très en vogue en ce début de XXI^e^ siècle en dehors des cas mentionnés *supra*, comme l'explique très clairement Venuti :

> *A translated text* [...] *is judged acceptable by most publishers, reviewers, and readers when it reads fluently, when the absence of any linguistic or stylistic peculiarities makes it seem transparent, giving the appearance that it reflects the foreign writer's personality or intention or the essential meaning of the foreign text—the appearance, in other words, that the translation is not in fact a translation, but the 'original'.* (1995 : 2)

15 Un bon traducteur serait alors comme un bon criminel : la traduction n'est réussie que s'il ne se fait pas prendre, c'est-à-dire si le texte traduit peut passer pour un texte écrit directement en langue cible sans être le fruit d'une traduction. Il cherche alors à atteindre ce que Salkie (2007) nomme en anglais la *naturalness* (le caractère naturel), qui va au-delà du grammaticalement correct et qui doit représenter aujourd'hui la priorité pour les traducteurs. C'est cette demande que l'on trouve sur le marché de la traduction à l'heure actuelle, tout au moins en ce qui concerne la traduction de qualité.

16 Afin d'atteindre ce but, le traducteur consulte souvent des documents rédigés en langue cible originale portant sur le même sujet que celui du texte à traduire, afin d'observer les choix terminologiques, phraséologiques et grammaticaux des experts du domaine. Au-delà de la consultation de quelques documents, il est possible de compiler un mini-corpus spécialisé maison (corpus DIY) afin de trouver des réponses précises à des problèmes de traduction que d'autres outils ne fourniraient pas. Cette méthode permet d'augmenter le

nombre de sources consultées, et donc de diminuer la prise de risque. Nous expliquons ci-dessous comment procéder à partir d'un problème de traduction précis.

Compilation et exploitation d'un mini-corpus spécialisé

17 Considérons le passage suivant, à traduire de l'anglais vers le français :

> *Real gross domestic product (GDP) increased 4.2 percent at an annual rate in the second quarter of 2014, according to the second estimate from the Bureau of Economic Analysis. The strong second-quarter growth represents a rebound from a first-quarter decline in GDP that largely reflected transitory factors like unusually severe winter weather and a sharp slowdown in inventory investment.* (publié par Jason Furman le 28 août 2014, < www.whitehouse.gov/blog/2014/08/28/second-estimate-gdp-second-quarter-2014 >)

18 Plusieurs questions de traduction liées à la nature spécialisée du texte pourront se poser ici au traducteur : comment traduire les termes économiques *annual rate*, *growth*, *rebound*, *decline*, *slowdown* ou encore l'adjectif *real* devant *GDP* ? Si l'utilisation de dictionnaires et de glossaires permet bien souvent de répondre à ces interrogations, dans le cas présent, la consultation de ces outils s'avère insuffisante. S'agissant de *annual rate*, les glossaires bien connus IATE et Termium Plus proposent tous deux *taux annuel* ; le site www.linguee.com amène une autre proposition, à savoir *rythme annuel*. Les termes désignant l'évolution à la hausse ou à la baisse du PIB ou de l'investissement (*growth*, *rebound*, *decline*, *slowdown*) ont quant à eux à chaque fois plusieurs traductions possibles, que ce soit en langue générale ou en langue spécialisée. Ainsi, pour ne prendre qu'un exemple, *decline* peut se traduire par *baisse*, *déclin*, *descente*, *chute*, *diminution*... ; la question se pose alors de savoir quel terme est compatible avec le contexte de l'article, ici en association avec le PIB (phénomène de collocation) : le PIB est-il en baisse, en déclin, en diminution... ? Convient-il d'employer plutôt un verbe : le PIB diminue, baisse, décroît, décline... ? Enfin, une traduction littérale de *real GDP*, à savoir *PIB réel*, est-elle possible, ou sommes-nous face à un phénomène de collocation non transposable en langue cible ?

19 Afin de répondre à ces questions, le traducteur a généralement tendance à effectuer une recherche documentaire, en consultant par exemple pour le cas qui nous concerne ici des articles journalistiques publiés sur le même sujet en langue cible originale (en français donc). Des articles publiés sur le même sujet (la situation de l'économie américaine aux deux premiers trimestres de 2014) et à la même période (été 2014) peuvent très facilement être trouvés en ligne :

« Le PIB américain revu à la hausse au deuxième trimestre à +4,6 % » (www.latribune.fr, 26/09/14)

« 4 % : le spectaculaire rebond de la croissance du PIB américain » (www.latribune.fr, 30/07/14)

« États-Unis : spectaculaire rebond de la croissance » (www.lemonde.fr, 30/07/14)

« États-Unis : le PIB encore révisé à la hausse » (www.lesaffaires.com, 26/09/14)

« Vif rebond de l'économie américaine au deuxième trimestre » (www.lesechos.fr, 30/07/14)

« La croissance américaine revue en hausse au 2e trimestre » (www.lexpress.fr, 26/09/14)

« Croissance de 4,6 % de l'économie américaine au second trimestre » (french.xinhuanet.com, 27/09/14)

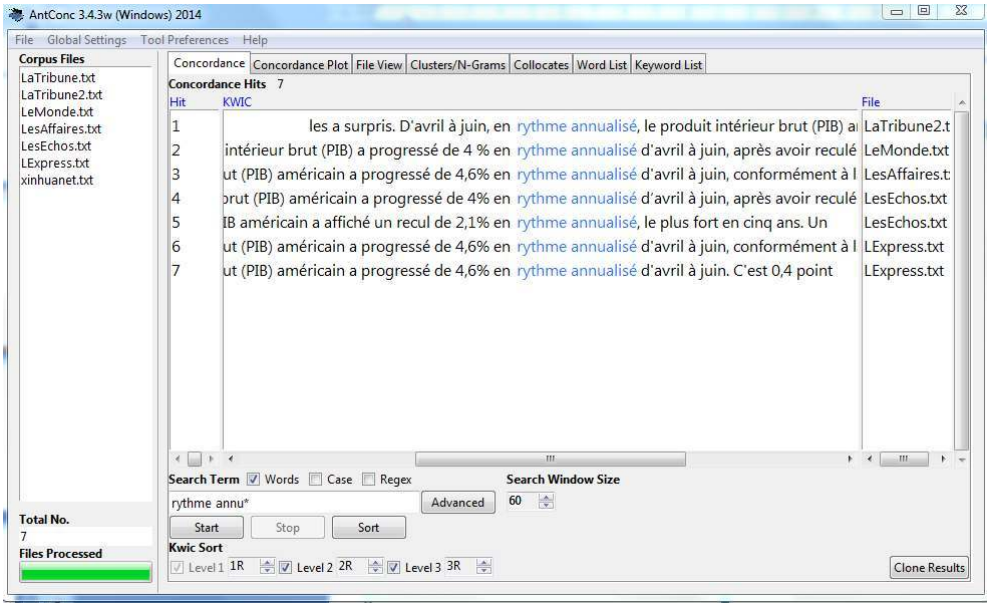
20 À partir de ces 7 articles rédigés en français original sur le même sujet, que nous pouvons supposer avoir été écrits par des spécialistes de la question, le traducteur peut alors, pour

plus de rapidité et d'efficacité, les rassembler en un mini-corpus spécialisé maison (corpus DIY) en sauvegardant chaque article en un fichier au format texte brut (.txt) avec encodage UTF-8 permettant la prise en charge des caractères spéciaux comme les caractères accentués. Il suffit pour cela de sélectionner grossièrement le texte sur la page internet en question, de le copier dans un fichier texte sans le nettoyer, c'est-à-dire sans retirer les photographies, légendes, publicités, hyperliens, etc. (il convient en revanche d'éviter d'inclure les commentaires d'internautes). Le fait de sauvegarder le fichier au format .txt permettra de ne conserver que le texte, qui certes inclura du texte « parasite » comme des légendes ou des publicités, mais celui-ci n'apparaîtra pas au moment de l'exploitation ciblée du corpus. Les fichiers n'ont donc pas besoin d'être « nettoyés », contrairement à ce qui se ferait en linguistique de corpus : nous prônons donc de conserver le fichier en l'état, et de recourir à un corpus qui n'aura pas été nettoyé et qui restera donc en partie « sale », ce qui n'aura aucune incidence sur les recherches effectuées (voir *infra*).

21 La compilation de ce corpus présente deux avantages : tout d'abord un gain de temps puisqu'il ne s'agit pas de *lire* l'ensemble des articles, mais d'*interroger* le corpus par le biais de mots-clefs (requêtes), mais aussi un gain de qualité puisque la consultation de 7 sources fournira une réponse plus sûre que la lecture de 2 ou 3 articles sur le sujet (nous ne sommes jamais à l'abri d'une qualité de langue non fiable : journaliste non francophone natif, non spécialiste du sujet, idiosyncrasies, ou tout simplement mauvais rédacteur). Il s'agit alors d'interroger le corpus, composé dans notre cas de 3 189 mots au total, afin de trouver rapidement la réponse à nos 3 questions avec une certaine confiance, grâce à un concordancier, à savoir un programme informatique qui va permettre de formuler des requêtes d'ordre linguistique afin d'observer le(s) terme(s) recherchés(s) dans leur contexte, en ayant un accès immédiat et visuellement pratique aux différentes occurrences du « pivot » (terme recherché) grâce au format d'affichage KWIC (*KeyWord In Context*) sous formes de lignes de concordance (voir captures d'écran *infra*). Nous utiliserons ici le concordancier AntConc, développé par Laurence Anthony et disponible gratuitement à l'adresse suivante : www.laurenceanthony.net/software/antconc/ (Anthony, 2014) ; nous n'exploiterons en revanche que quelques fonctionnalités et ne proposerons pas un descriptif exhaustif de tout ce que l'outil peut faire.

22 S'agissant de la traduction de *annual rate*, le traducteur pourra tester la pertinence des traductions proposées par les glossaires en ligne ou par le site Linguee, pour s'apercevoir que la requête « rythme annuel » ne donne aucune réponse positive, ce qui, sans exclure totalement cette possibilité (cf. modeste taille du corpus et le principe général qui dit que ce n'est pas parce que l'on ne trouve pas une expression dans un corpus que cela signifie qu'elle n'existe pas ; on parle alors de « silence », dû à la composition du corpus ou à la requête trop restrictive), jette toutefois le doute sur cette traduction de *annual rate*. En revanche, si l'on souhaite tester une hypothèse et entrer la requête « rythme annu* » afin de vérifier si un autre adjectif dérivé de *année* est possible, ceci donne 7 résultats positifs, tous pour des occurrences de *rythme annualisé* (figure 2). Une lecture des lignes de concordance montre que le sens de *rythme annualisé* correspond au sens de *annual rate*. Le traducteur peut à cette occasion récupérer une autre information cruciale : le cotexte linguistique au sein duquel *rythme annualisé* apparaît. Les 7 occurrences font état de l'utilisation systématique de la préposition *en* au sein d'un schéma récurrent : « x % en rythme annualisé (de x à x) ». Cette rapide requête donne donc des informations phraséologiques cruciales pour un emploi fluide de l'expression *rythme annualisé*.

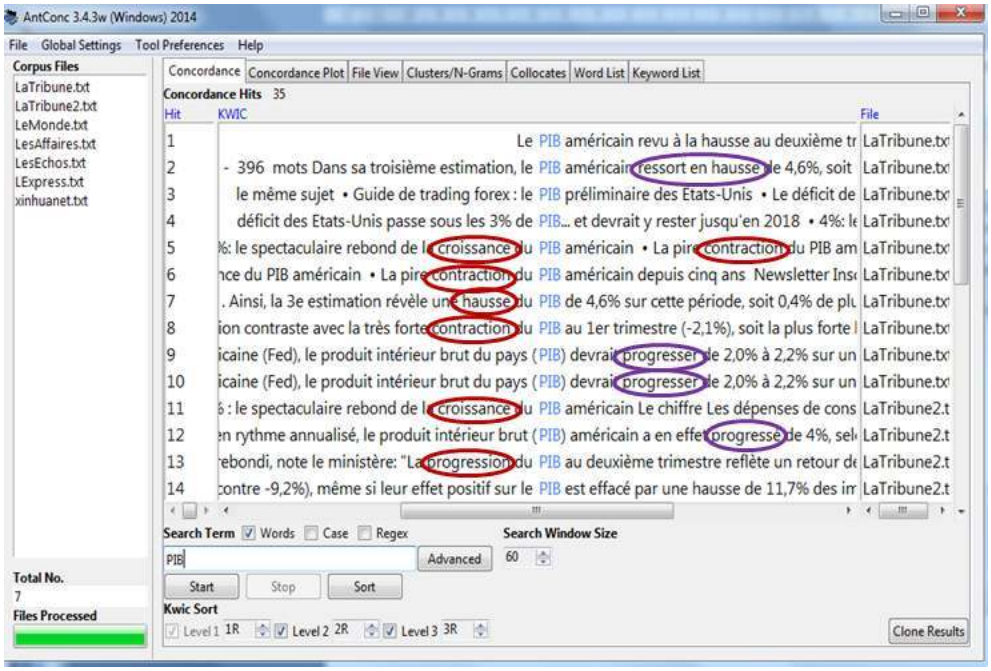
Figure 2. – Résultats de la requête « rythme annu* ».



23 Selon cette méthodologie, il s'agit donc de *tester des hypothèses* et de vérifier si une requête donne des résultats ou non ; il n'est pas possible d'interroger le corpus directement afin d'obtenir une réponse à une question de traduction du type « comment traduire *annual rate* ? ». Par ailleurs, ce type de corpus n'est pas étiqueté et est composé de texte brut uniquement[1] ; il n'est donc pas possible d'effectuer de recherches complexes intégrant des codes grammaticaux (exemple : quels adjectifs trouve-t-on après *PIB* ?) comme cela est possible dans les corpus électroniques de référence comme Frantext, mais l'absence d'étiquetage permet une compilation rapide et non technique du corpus de travail.

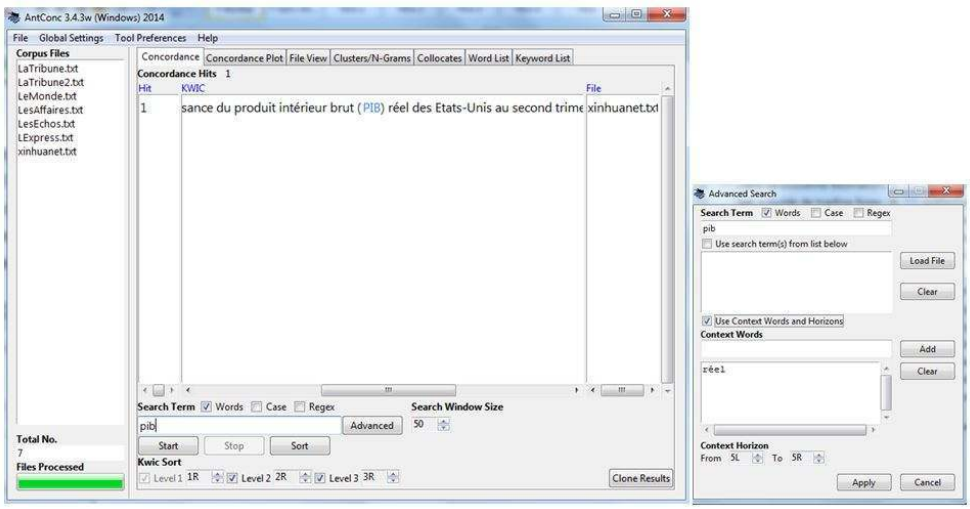
24 S'agissant de la traduction des termes renvoyant à l'évolution à la hausse ou à la baisse du PIB (*growth*, *rebound*, *decline*), qui pose en fait la question du type de verbe ou de nom pouvant s'associer en français économique original avec le PIB d'un pays (phénomène de collocation), il est possible d'effectuer une requête sur le terme « PIB » et d'observer les cotextes droit et gauche, observation qui révèle les noms (*croissance*, *contraction*, *hausse*, *progression*, *recul*, *expansion*) et les verbes (*ressortir en hausse*, *progresser*, *passer au-dessus/en-dessous*, *reculer*, *afficher un recul*) apparaissant dans le cotexte immédiat du sigle *PIB* (voir la figure 3 pour les premiers exemples, la mise en relief des termes ayant été effectuée par nos soins). À partir de ces résultats, il semble que *decline* pourra se traduire par *recul*, *baisse* ou encore *contraction* et *growth* par *croissance* ou *progression*, sachant que l'emploi de verbes par recatégorisation est également possible.

Figure 3. – Résultats de la requête « PIB ».



25 Enfin, s'agissant de notre troisième question (comment traduire *real GDP* ?), il est possible de vérifier dans le corpus l'existence de cette cooccurrence en français original, grâce à la requête « PIB » combinée à une recherche de contexte dans AntConc (recherche avancée ['Advanced'], en association avec *réel* à plus ou moins 5 mots à droite ou à gauche du pivot). Le corpus contient un résultat de ce type (figure 4) correspondant à cette requête avancée, ce qui atteste de l'existence de cette expression en français original, *contra* des requêtes concernant *véritable* et *vrai*, qui ne fournissent aucun résultat. Une fois encore, il s'agit de tester des hypothèses.

Figure 4. – Résultats de la requête avancée « PIB + réel ».



26 Grâce à ce mini-corpus DIY spécialisé en langue cible, il nous a donc été possible de trouver des réponses que d'autres outils ne fournissaient pas. La procédure n'est ni technique ni chronophage (le seul aspect technique étant peut-être la sauvegarde en

fichier .txt encodage UTF-8, mais ceci est très simple, quel que soit le logiciel de traitement de texte utilisé), et la prise de risques est réduite. Le gain en qualité est donc évident. S'agissant de la productivité, il est possible d'automatiser en grande partie la procédure de compilation du corpus grâce à un logiciel spécifique, en l'occurrence BootCaT (*Bootstrap Corpora And Terms from the Web*), qui est un logiciel frontal (*front-end*) permettant de compiler un corpus de façon automatique à partir de mots-clefs (*seeds*) choisis par l'utilisateur ; le logiciel récupère alors sur l'internet des pages web correspondant à ces mots-clefs (Baroni & Bernardini, 2004). Le corpus peut ensuite être exploité à l'aide du concordancier AntConc comme nous l'avons fait *supra* pour le corpus compilé manuellement. Si la qualité des textes récupérés automatiquement peut être moins fiable (il est néanmoins possible d'appliquer des filtres et de contrôler les documents sélectionnés), la quantité plus importante de données ainsi collectées permet de compenser l'absence de vérification manuelle. C'est ce que montre en tout cas notre expérience en la matière (voir Loock, à paraître, pour une explication et application de la procédure).

Compilation d'un corpus comparable

De la grammaire comparée à la traductologie : quels enjeux pour le traducteur ?

27 La traduction entre deux langues cousines, l'une romane et l'autre germanique, peut parfois donner l'illusion de similitudes parfaites, d'équivalences directes entre deux termes dits « transparents » (p. ex. les lexèmes *thing* et *chose*), entre deux structures syntaxiques (p. ex. les structures existentielles en *there is/are* et en *il y a*), ou encore pour le même type de phrase (p. ex. les phrases à la voix passive). Or, comme l'ont montré de nombreux ouvrages comparatistes (p. ex. Chuquet & Paillard, 1987 ; Guillemin-Flescher, 1981 ; Vinay & Darbelnet, 1958), l'équivalence n'est jamais totale. Ainsi, en dehors des contraintes purement syntaxiques (p. ex. non-existence du passif prépositionnel ou du passif datif en français), la voix passive est bien plus fréquente en anglais qu'elle ne l'est en français, ce qui doit avoir des conséquences au moment de la traduction : une traduction systématique de phrases en anglais à la voix passive par des phrases en français à la voix passive risque de donner au texte traduit une forme d'artificialité, que certains qualifieront de calques maladroits, loin de l'invisibilité souvent recherchée. Il est donc important que le traducteur connaisse ces différences de fonctionnement et de distribution entre les deux langues, en particulier lorsqu'il s'agit d'équivalents traductionnels. Le traducteur doit alors compter sur sa fine connaissance des deux langues dans leur usage, et donc sur son intuition, elle-même fondée sur son expérience, ou encore sur les différents ouvrages comparatistes mentionnés *supra*. Mais il est également possible de mettre au jour ces différences inter-langagières à l'aide de corpus électroniques, que ceux-ci soient les corpus de référence comme le BNC, le COCA ou Frantext pour l'anglais et le français respectivement, ou en compilant ses propres corpus de langues originales lorsque les corpus de référence ne suffisent pas. Il s'agit alors de recourir à un corpus comparable, puisqu'il s'agit ici de comparer deux sous-échantillons, dont on s'assurera de la comparabilité au sens technique du terme (similitude au niveau des types de textes, des registres, des périodes, des modes de production), afin de mettre

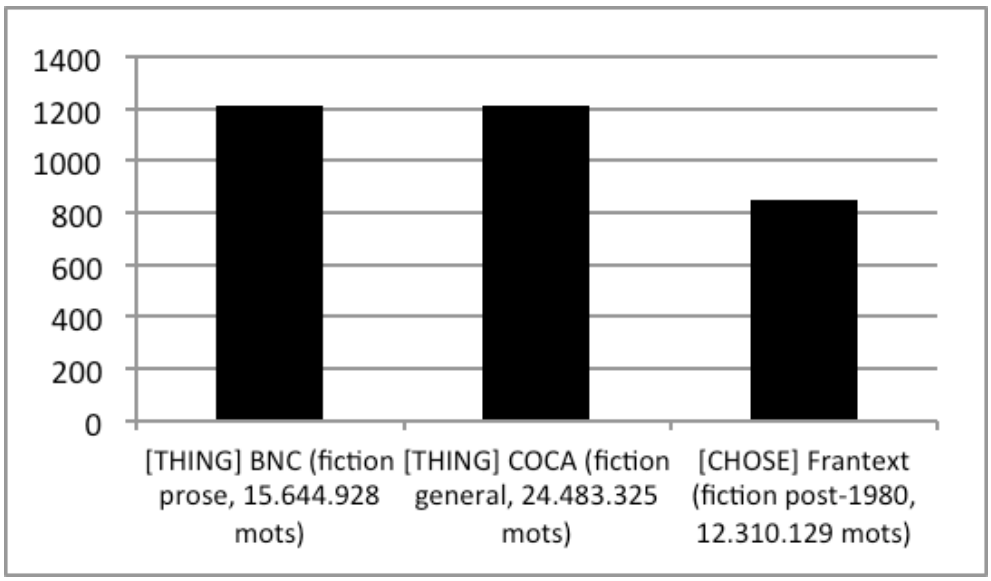
au jour similitudes et différences d'usage qui correspondent à des normes venant s'ajouter aux règles morphosyntaxiques.

28 Il conviendra alors de prendre en compte cet usage de la langue au moment de la traduction, ce que font les locuteurs natifs lorsqu'ils s'expriment en langue originale de façon totalement inconsciente mais qui posera problème au traducteur qui risque d'être influencé par la langue source et de sélectionner un équivalent traductionnel sans prendre en compte l'usage de la langue cible. Ce type de différences relevant de l'usage de la langue pose problème dans la mesure où il n'est pas codifié, c'est-à-dire recensé par les dictionnaires ou les ouvrages de grammaire. C'est alors par le biais de comparaisons inter-langagières, entre deux échantillons de langues originales, qu'il est possible de les mettre au jour, en les quantifiant.

Études à partir de corpus existants

29 La mise en évidence de différences inter-langagières (ici entre l'anglais et le français) peut se faire à partir de corpus existants comme le BNC ou le COCA pour l'anglais et Frantext pour le français. Il s'agira alors de comparer la fréquence d'un phénomène linguistique (mot, type de mot, structure...) dans les deux langues. Ainsi, il sera possible, de façon très simple et très rapide, de mettre au jour une différence d'usage importante pour les lemmes [THING] *vs* [CHOSE], considérés pourtant comme des équivalents traductionnels, en quantifiant leur fréquence au sein d'échantillons comparables (textes de fiction publiés après 1980), extraits des trois corpus de référence mentionnés *supra* (figure 5). Les résultats montrent que [THING] est bien plus fréquent en anglais littéraire (britannique ou américain) que ne l'est [CHOSE] en français littéraire, et ce de façon extrêmement significative ($p<0.0001$)[2].

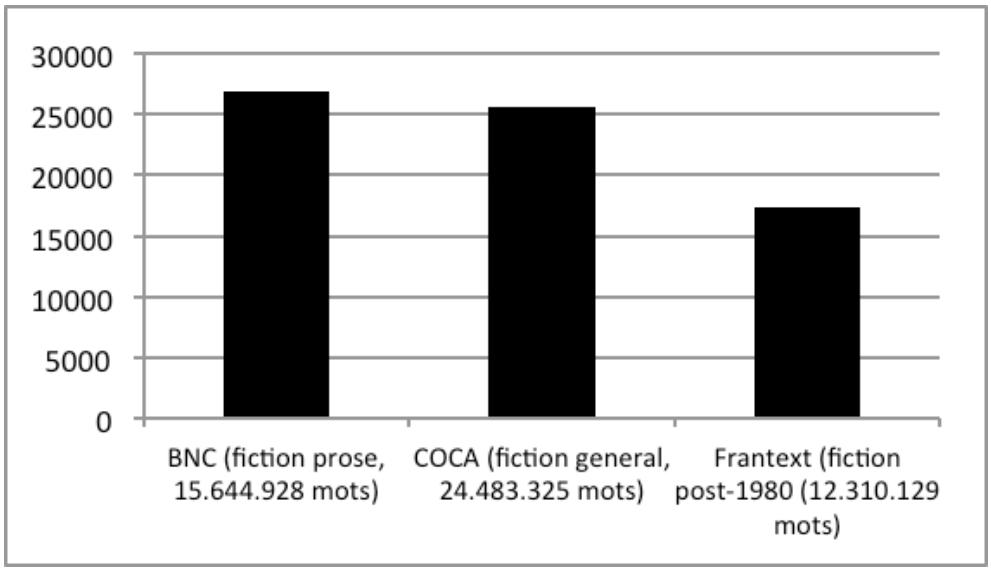
Figure 5. – Fréquence (par million de mots) des lexèmes [CHOSE] et [THING] dans les textes de fiction en anglais et en français.



30 Selon une méthode similaire, il sera possible d'établir que le coordonnant *and* est bien plus fréquent en anglais (britannique/américain) que le coordonnant *et* ne l'est en

français, et ce de façon extrêmement significative (p<0.0001), comme le montre la figure 6.

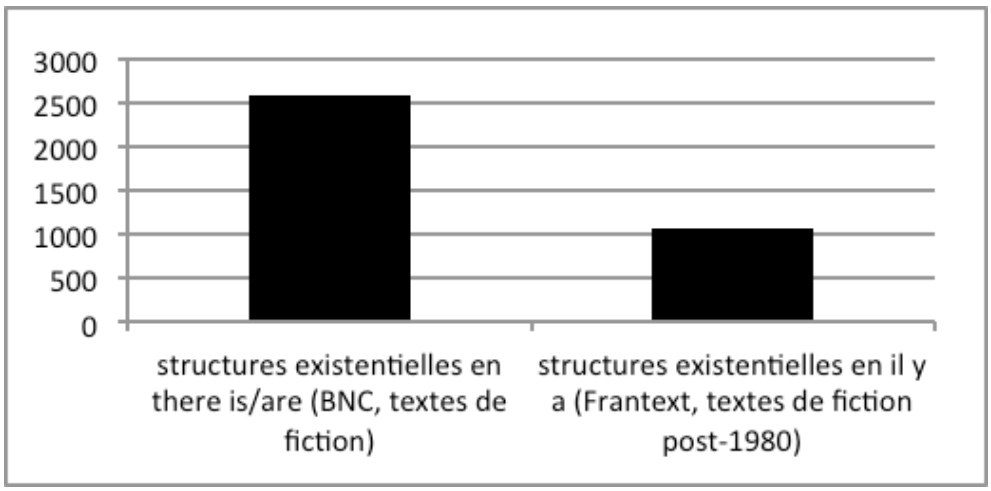
Figure 6. – Fréquence (par million de mots) des coordonnants *and* et *et* dans les textes de fiction en anglais et en français.



31 De la même manière, on pourra s'intéresser à une structure syntaxique et montrer, comme dans Cappelle et Loock (2013), que toutes choses étant égales par ailleurs, les structures existentielles, à savoir les structures introduites par *there* en anglais et les structures en *il y a* en français (1)-(2), bien qu'elles soient équivalentes d'un point de vue traductionnel, sont bien moins fréquentes en français qu'elles ne le sont en anglais (figure 7)[3].

(1) *There is of course no logical reason why things should be different this time*
(2) Sans réfléchir, je réponds que, dans cette armoire, il y a un émetteur-récepteur avec lequel j'envoie chaque soir des messages codés à Londres.

Figure 7. – Fréquence (par million de mots) des structures existentielles dans les textes de fiction en anglais et en français.



Études à partir d'un corpus DIY : l'exemple des textes nécrologiques

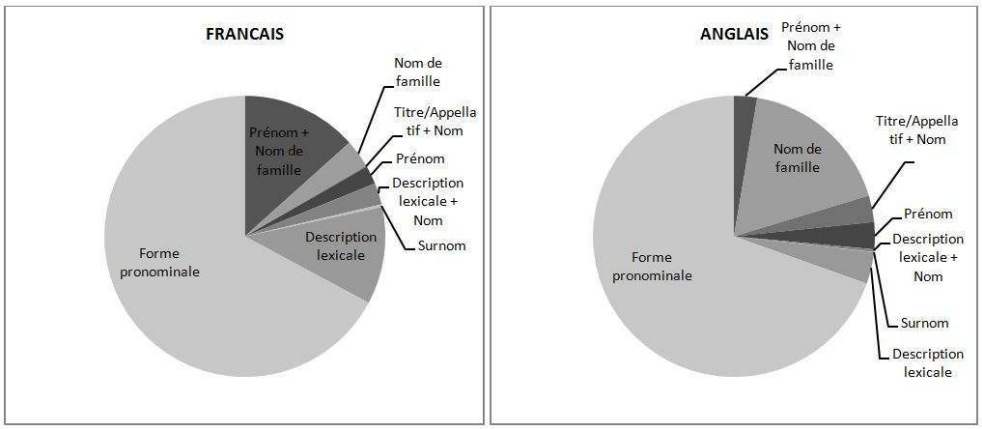
32 Lorsque les corpus existants ne suffisent pas, par exemple lorsque le traducteur est confronté à un type de textes très spécialisés, il est possible de compiler son propre corpus. Ceci prend davantage de temps que la compilation d'un mini-corpus spécialisé en langue cible (voir *supra*), mais si ce type de projets est récurrent pour le traducteur, alors le rapport coût/bénéfice sera intéressant ; dans tous les cas, le gain de qualité est certain. Nous prenons ici l'exemple des articles nécrologiques publiés dans les médias lors du décès d'une célébrité, qui sont des textes très codifiés du point de vue de leur structure (annonce de la mort, date, lieu et circonstances de l'événement, exposé plus ou moins long sur sa vie et sa carrière) mais aussi de leurs caractéristiques linguistiques, faisant de ce type de texte un « minilecte », doté d'un jargon spécifique, d'une syntaxe codifiée, de conventions discursives précises, soit tout un ensemble de phénomènes linguistiques qui vont au-delà de la bonne terminologie et du grammaticalement correct (voir Nordman, 1996 : 556 pour une définition). Dans ce type de texte, il est évidemment très souvent fait référence à la personne décédée, au moyen de différents procédés linguistiques, que ce soit en anglais ou en français. Ainsi, dans une nécrologie dédiée à Michael Jackson, il sera possible de référer à l'artiste de différentes manières :

1. Prénom (+ deuxième prénom) + Nom de famille : Michael (Joseph) Jackson
2. Nom de famille : Jackson
3. Prénom : Michael
4. Description lexicale (+ Prénom) + Nom de famille : American singer Michael Jackson, le chanteur américain Michael Jackson
5. Titre/Appellatif + Nom de famille : Mr/M. Jackson
6. Surnom : Bambi, Wacko Jacko
7. Description lexicale : the American singer/le chanteur américain
8. Formes pronominales

33 La grammaire des deux langues offrant les mêmes possibilités, il pourra être intéressant de comparer la fréquence d'utilisation de ces différentes formes de référence à la personne. Pour cela, nous avons compilé un corpus comparable de nécrologies rédigées en français original et en anglais original, baptisé « Nécrocorpus » (Loock & Lefebvre-Scodeller, 2014) et contenant des textes nécrologiques issus de la presse française (*Libération*, *Le Figaro*, *Le Monde*, *Le Point*...) et de la presse anglaise/américaine (*The Times*, *The New York Times*, *The Independent*, *The Guardian*...). Ce corpus a contenu dans un premier temps (étude pilote) 100 nécrologies dans les deux langues (50 en anglais, 50 en français) relatives à quatre célébrités décédées (Michael Jackson, Amy Winehouse, Larry Hagman, et Whitney Houston) avant d'être complété par des étudiants en formation avancée avec 100 nécrologies supplémentaires pour d'autres célébrités. Nous avons comptabilisé dans chaque sous-corpus le nombre d'occurrences de chaque type de référence à la personne décédée, en prenant bien soin d'omettre les citations (rapportées et non produites par le journaliste). La figure 8 montre l'existence de similitudes et de différences inter-langagières de fréquence : d'un côté, on note que la répartition entre les formes lexicales (1-7) et les formes pronominales (8) est la même dans les deux langues, tout comme la fréquence d'emploi des prénoms ou encore des surnoms (très marginaux) ; de l'autre, on constate que certains types d'expressions référentielles sont plus fréquemment utilisés dans une langue que dans l'autre : ainsi le nom de famille seul et l'emploi d'un appellatif ont une fréquence plus élevée dans les nécrologies écrites en anglais que dans celles

écrites en français, tandis que dans ces dernières on y trouve davantage de descriptions lexicales, accompagnées ou non du nom de la célébrité, et davantage de noms complets (prénom suivi du nom de famille).

Figure 8. – Comparaison anglais original-français original pour les différentes façons de référer à la personne dans des textes nécrologiques (étude pilote).



34 À partir de ces résultats, nous avons souhaité mener une expérience avec des étudiants au sein d'une formation pour futurs traducteurs spécialisés, en l'occurrence le Master « Traduction Spécialisée Multilingue » (TSM) de l'Université de Lille. Il a été demandé dans un premier temps aux étudiants en début d'année universitaire de traduire un texte nécrologique, relatif au décès de l'acteur américain Cory Monteith, de l'anglais vers le français. Dans un deuxième temps, il leur a été demandé de mener une étude comparatiste similaire à l'étude que nous venons d'évoquer, et éventuellement dans un troisième temps de post-éditer leur traduction à la lumière des résultats obtenus (Loock, 2014, 2015).

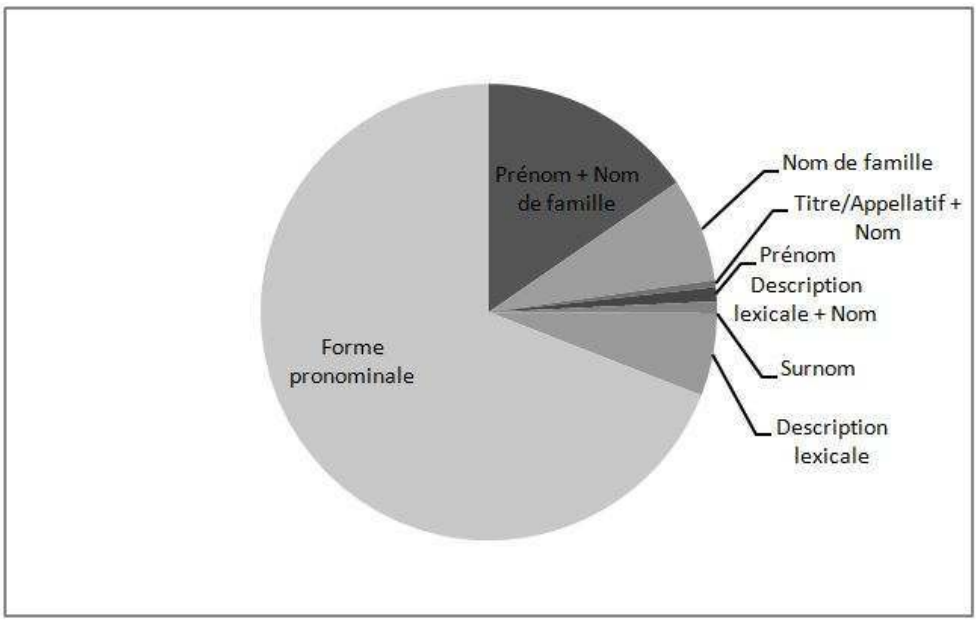
35 Avant l'étude de corpus, les traductions de ces étudiants avancés faisaient état de certains « calques » en ce qui concerne les expressions de référence à la personne. Ainsi, on pouvait noter notamment ce que nous considérons être une sur-utilisation du nom de famille employé seul, ce qui donnait parfois aux traductions un caractère quelque peu artificiel :

> (3) Le corps de Monteith a été découvert au Fairmont Pacific Rim Hotel un peu après midi.
> (4) Monteith, qui est né à Calgary, Alberta et a grandi à Victoria en Colombie Britannique, était le plus jeune fils de Joe Monteith, militaire de l'Infanterie légère canadienne de la princesse Patricia (PPCLI) et d'Ann McGregor, décoratrice d'intérieur. Ses parents divorcèrent lorsqu'il était très jeune, et Monteith eut beaucoup de mal à s'adapter au bouleversement de sa vie de famille. En conséquence de ces difficultés, Monteith fréquenta 16 écoles différentes — dont une pour adolescents perturbés — et abusa des drogues et de l'alcool.

36 À l'inverse, les étudiants n'avaient eu que très peu recours, lors de cette première traduction « naïve » (ils n'étaient pas au courant de l'objectif final de ce travail et de l'étude sur corpus qui devait suivre), aux descriptions lexicales (*l'acteur américain*, *l'acteur de la série américaine* Glee, *le compagnon de l'actrice Lea Michele*), mais ils avaient eu davantage recours au nom complet (*Cory Monteith*). La figure 9 fournit les résultats de

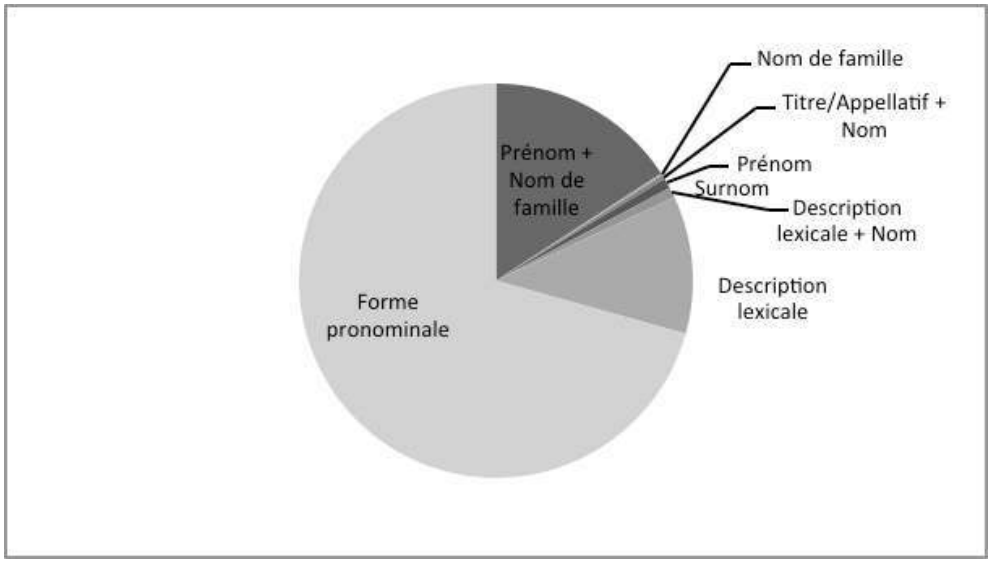
l'analyse de ces premières traductions (16 au total, tous les étudiants étant francophones de naissance).

Figure 9. – Résultats première traduction (« naïve »).



37 Après cette traduction effectuée avant le début des enseignements, les étudiants ont été amenés à compiler leur propre corpus et à mesurer la fréquence des différentes expressions de référence à la personne décédée en anglais et en français, selon la même méthodologie que celle utilisée par Loock et Lefebvre-Scodeller (2014). Ils ont alors compilé un nouveau corpus de 100 nécrologies en anglais original et en français original, relatives aux décès de cinq nouvelles célébrités (Margaret Thatcher, Jean-Paul II, Heath Ledger, Elisabeth Taylor, Steve Jobs). En dehors de quelques données plus complexes[4], les étudiants ont obtenu les mêmes résultats que pour les quatre premières célébrités. À la lumière de ceux-ci, ils ont alors eu la possibilité, sans en avoir toutefois l'obligation, de post-éditer leur première traduction. Tous les étudiants ont opté pour cette possibilité et ont souhaité apporter des modifications à leur traduction initiale. Par exemple, conscients du fait que l'utilisation du nom propre seul (*Monteith*) était marginale en français, alors que l'on en trouve 10 occurrences dans le texte anglais original, les étudiants ont remplacé cette utilisation du nom propre seul qu'ils avaient eu tendance à calquer lors de leur première traduction. De la même manière, ils ont diminué le nombre d'occurrences du nom complet (*Cory Monteith*) lorsque cette possibilité avait été systématiquement utilisée en lieu et place du nom propre seul. À l'inverse, le nombre de descriptions lexicales (p. ex. *l'un des acteurs principaux de la série* Glee) et de formes pronominales (p. ex. *il*, *lui*) a subi une augmentation. La figure 10 fournit les résultats de l'analyse des traductions post-éditées. Les étudiants ont par ailleurs effectué d'autres modifications suite à l'observation de leur corpus : remplacement du passé simple par le passé composé, nominalisations (p. ex. *Lorsque Cory a 19 ans* > *À l'âge de 19 ans*), modifications lexicales (p. ex. *décédé* > *retrouvé mort* ; *tristesse* > *chagrin*).

Figure 10. – Résultats traduction post-éditée.



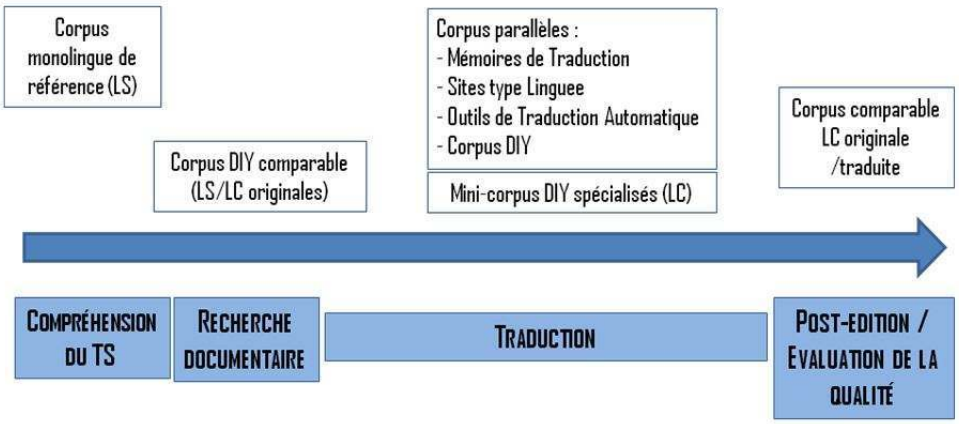
38 La compilation d'un corpus comparable a donc permis aux étudiants de prendre conscience des différences d'usage entre les deux langues et les a amenés à modifier leur traduction initiale. En revanche, il convient de noter que les secondes traductions font la part belle aux stratégies les plus utilisées (forme pronominale, prénom + nom, description lexicale) et semblent laisser de côté des possibilités plus marginales. Il y a donc là un risque de standardisation de la langue cible, à laquelle il convient de sensibiliser les étudiants.

Discussion et conclusions

39 L'objectif de cet article était de montrer que les corpus électroniques (en dehors de ceux intégrés aux logiciels de TAO, aux outils de TA, et au sein de certains sites internet) pouvaient être considérés comme des outils de TAO : ils assistent en effet le traducteur dans son travail et sont compilables/exploitables à partir d'un ordinateur. Ils peuvent fournir des réponses précises à des problèmes de traduction précis que d'autres outils ne permettent pas toujours de résoudre. Face aux problèmes ergonomiques réels, il est important de développer un protocole de compilation et d'exploitation simple et rapide, et nous avons formulé des propositions en ce sens, l'objectif final étant d'accroître la fluidité du texte cible, et donc la qualité de la traduction. En particulier, nous avons soutenu l'idée que les approches du linguiste chercheur et du traducteur ne devaient pas être confondues : si le premier observe la langue afin d'en tirer des conclusions selon une approche scientifique qui exige une certaine rigueur, le second peut se permettre quelques raccourcis (ni nettoyage du corpus ni étiquetage). De la même manière, nous avons montré que des corpus de taille tout à fait modeste comme les mini-corpus spécialisés en langue cible originale fournissaient des résultats pertinents, sans qu'il faille recourir aux méga-corpus qui contiennent des centaines de millions, voire des milliards de mots. Il s'agit donc de faire simple, en considérant les objectifs du traducteur professionnel, qui ne sont pas ceux du linguiste chercheur, sans toutefois engendrer des problèmes de qualité bien entendu (il ne s'agit pas de collecter des textes à l'aveugle non plus).

40 Également, il importe de bien comprendre que les différents types de corpus que nous avons mentionnés dans ce chapitre ont une utilité qui varie en fonction du type de texte à traduire mais surtout du type d'informations recherchées, et donc de l'étape du projet de traduction concernée. Ainsi si l'on ne traduit qu'une seule fois un bulletin météo, un mini-corpus spécialisé de bulletins météo en langue cible pourra suffire ; s'il s'agit d'un type de projets récurrent, la compilation d'un corpus comparable langue source/langue cible originales pourra s'avérer rentable dans le cadre de la phase de recherche documentaire (voir notre exemple sur les textes nécrologiques). Les corpus monolingues de référence en langue source, eux, seront utiles pour la phase de compréhension du texte source dans ses moindres détails tandis que les corpus parallèles (mémoires de traduction, sites spécialisés, outils de traduction automatique) seront principalement utiles pendant la phase de traduction elle-même, au moment de la prise de décision, en tant que sources d'inspiration pour le traducteur. Enfin, un type de corpus que nous n'avons pas eu la place d'aborder ici pourra s'avérer utile dans le cadre de la post-édition et de l'évaluation de la qualité : il s'agit du corpus comparable langue originale/langue traduite pour la même langue, en l'occurrence la langue cible, afin de vérifier l'homogénéisation linguistique (qui selon nous va de pair avec l'invisibilité du traducteur) entre le texte traduit et la langue originale (Loock, 2012b ; Loock *et al.*, 2014). La figure 11, sur laquelle nous concluons cet article, illustre de façon synthétique le lien entre les différentes étapes du processus de traduction et le type de corpus à utiliser.

Figure 11. – L'utilisation des différents types de corpus électroniques selon l'étape du processus de traduction (TS = texte source ; LS = langue source ; LC = langue cible).



BIBLIOGRAPHIE

Ahmad Khurshid, Davies Andrea, Fulford Heather & Rogers Margaret (1994), « What's in a Term? The Semi-automatic Extraction of Terms from Text », M. F. Pochhacker & K. Kaind (dir.), *Translation Studies. An Inter-discipline*, Amsterdam/Philadelphie : John Benjamins, 267-278.

ANTHONY Laurence (2014), *AntConc* (Version 3.4.3) [Computer Software], Tokyo, Japon : Waseda University, en ligne sur <www.laurenceanthony.net>.

BAKER Mona (1992), *In Other Words: A Coursebook of Translation*, Londres : Routledge.

BARONI Marco & BERNARDINI Silvia (2004), « *BootCaT*: Bootstrapping corpora and terms from the web », *actes du colloque LREC 2004*, <www.lrec-conf.org/proceedings/lrec2004/>.

BEEBY Alison, RODRIGUEZ Ines Patricia, & SANCHEZ-GIJON Pilar (dir.) (2009), *Corpus Use and Translating*, Amsterdam/Philadelphie : John Benjamins.

BOULTON Alex (2008), « Esprit de corpus : promouvoir l'exploitation de corpus en apprentissage des langues », *Texte et Corpus*, *3*, 37-46.

BOULTON Alex (2012), « Beyond concordancing: Multiple affordances of corpora in university language degrees », *Procedia – Social and Behavioral Sciences*, *34*, 33-38.

BOWKER Lynne (2006), « Investigating Semantic Prosody in Language for Special Purposes: A Corpus-based Study in the Specialized Field of Labour Relations with some Implications for Translation », *Translation Studies in the New Millennium*, *4*, 15-31.

BOWKER Lynne & PEARSON Jennifer (2002), *Working with specialized language: a practical guide to using corpora*, Londres/New York : Routledge.

BOWKER Mickael & BARLOW Lynne (2008), « A comparative evaluation of bilingual concordancers and translation memory systems », E. Yuste Rodrigo (dir.), *Topics for Language Resources in Translation and Localisation*, Amsterdam/Philadelphie : John Benjamins, 1-22.

CAPPELLE Bert & LOOCK Rudy (2013), « Is there Interference of Usage Constraints? A Frequency Study of Existential *there is* and its French Equivalent *il y a* in Translated *vs* Non-Translated texts », *Target*, *25*(2), 252-275.

CHUQUET Hélène & PAILLARD Michel (1987), *Approche linguistique des problèmes de traduction anglais-français*, Paris : Ophrys.

DAVIES Mark (2004-), *BYU-BNC* (based on the British National Corpus from Oxford University Press), en ligne sur <corpus.byu.edu/bnc/>.

DAVIES Mark (2008-), *The Corpus of Contemporary American English: 450 million words, 1990-present*, en ligne sur <corpus.byu.edu/coca/>.

DAVIES Mark (2010-), *The Corpus of Historical American English: 400 million words, 1810-2009*, en ligne sur <corpus.byu.edu/coha/>.

FRÉROT Cécile (2010), « Outils d'aide à la traduction : pour une intégration des corpus et des outils d'analyse de corpus et dans l'enseignement de la traduction et la formation des traducteurs », *Les Cahiers du GEPE* 2, <www.cahiersdugepe.fr/index.php?id=1164>.

GUILLEMIN-FLESCHER Jacqueline (1981), *Syntaxe comparée du français et de l'anglais. Problèmes de traduction*, Gap/Paris : Ophrys.

JOHANSSON Stig & OKSEFJELL Signe (dir.) (1998), *Corpora and cross-linguistic research: theory, method and case studies*, Amsterdam/New York : Rodopi.

KÜBLER Natalie (2001), « Corpora in Terminology and Translation teaching: methodological approach », S. de Cock, G. Gilquin, S. Granger, & S. Petch-Tyson (dir.), *Proceedings of the ICAME 01 Conference*, Louvain-la-Neuve, 53-55.

KÜBLER Natalie (2008), « A Comparable Learner Translator Corpus: creation and use », P. Zweigenbaum (dir.), *Proceedings of the Comparable Corpora Workshop of the LREC Conference*, Marrakech, 73-78.

KÜBLER Natalie (2011), « Working with different corpora in translation teaching », A. Frankenberg-Garcia, L. Flowerdew & G. Aston (dir.), *New Trends in Corpora and Language Learning* , Londres : Continuum, 62-80.

KÜBLER Natalie & ASTON Guy (2010), « Using Corpora in Translation », M. Mc Carthy & A. O'Keefe (dir.), *The Routledge Handbook of Corpus Linguistics*, Londres : Routledge, 505-515.

KÜBLER Natalie & VOLANSCHI Alexandra (2012), « Semantic prosody and specialised translation, or how a lexico-grammatical theory of language can help with specialised translation », A. Boulton, S. Carter-Thomas, & E. Rowley-Jolivet (dir.), *Corpus-informed Research and Learning in ESP : Issues and Applications*, Amsterdam/Philadelphie : John Benjamins, 105-135.

KUNZ Kerstin, CASTAGNOLI Sara & KÜBLER Natalie (2010), « Corpora in translator training: A program for an eLearning course », D. Gile, H. Gyde, & N. K. Pokorn (dir.), *Why Translation Studies Matters*, Amsterdam/Philadelphie : John Benjamins, 195-208.

LOOCK Rudy (2012a), « La traductologie de corpus : étude de cas et enjeux », N. D'Amelio (dir.), *Au cœur de la démarche traductive : débat entre concepts et sujets, actes du colloque international « Traduction / Traductologie. Conceptualisations et nœuds de subjectivité en traduction »*, Mons : CIPA, 99-116.

LOOCK Rudy (2012b, octobre), « Doit-il/Peut-il y avoir homogénéisation entre langue originale et langue traduite ? », communication présentée au colloque *La Cohérence discursive à l'épreuve : traduction et homogénéisation*, Paris, France.

LOOCK Rudy (2014), « Using corpora in the classroom to improve translation quality: report on an experiment », communication présentée à la journée d'études *Traduction et Qualité*, Lille, France.

LOOCK Rudy (2015), « From comparative grammar to translation studies: the use of DIY comparable corpora in the translation classroom », communication présentée au colloque *engcorpora2015*, Paris, France, <www.youtube.com/watch ?v =kFclzdePyyA>.

LOOCK Rudy (à paraître), *La Traductologie de corpus*, Lille : Presses Universitaires du Septentrion.

LOOCK Rudy, MARIAULE Michaël & OSTER Corinne (2013), « Traductologie de corpus et qualité : étude de cas », *actes du colloque Tralogy II, session 5 – Assessing Quality in MT/Mesure de la qualité en TA*, <lodel.irevues.inist.fr/tralogy/index.php?id=243>.

LOOCK Rudy & LEFEBVRE-SCODELLER Cindy (2014), « Writing about the Dead: a Corpus-based Study on how to Refer to the Deceased in French *vs* English obituaries », *Current Trends in Translation Teaching and Learning E*, *1*, <www.cttl.org/cttl-e-2014.html>.

NORDMAN Marianne (1996), « Cooking recipes and knitting patterns: Two minilects representing technical writing », K. D. Baumann & H. Kalverkämper (dir.), *Fachliche Textsorten: Komponenten-Relationen-Strategien*, Tübingen : Gunter Narr Verlag, 554-575.

OLOHAN Maeve (2004), *Introducing corpora in translation studies*, Londres : Routledge.

SALKIE Raphael (1997), « Naturalness and contrastive linguistics », J. Melia & B. Lewandowska-Tomaszczyk (dir), *Proceedings of PALC 97*, Lodz : University of Lodz, 297-312, reproduit dans W. Teubert & R. Krishnamurthy (dir.), *Corpus Linguistics (Critical Concepts in Linguistics)*, *4*, Londres : Routledge, 336-351.

TENGKU Mahadi Tengku Sepora, VAEZIAN Helia, & AKBARI Mahmoud (2010), *Corpora in translation : a pratical guide*, Bern : Peter Lang.

VARANTOLA Krista (2003), « Translators and disposable corpora », F. Zanettin, S. Bernard, & D. Stewart (dir.), *Corpora in Translator Education*, Manchester/Northampton : St Jerome, 55-70.

VENUTI Lawrence (1995), *The Translator's Invisibility*, New York : Routledge.

VINAY Jean-Paul & DARBELNET Jean (1958), *Stylistique comparée du français et de l'anglais*, Paris : Didier.

ZANETTIN Federico (2012), *Translation-driven corpora*, Manchester : St Jerome Publishing.

ZANETTIN Federico, BERNARDINI Silvia & STEWART Dominique (dir.) (2003), *Corpora in Translator Education*, Manchester/Northampton : St Jerome Publishing.

NOTES

1. À noter qu'il est possible d'exploiter des corpus étiquetés au sein du concordancier, à partir de fichiers préalablement annotés au moyen d'étiqueteurs automatiques comme TreeTagger par exemple ou encore du logiciel TagAnt, développé également par Laurence Anthony. Nous n'entrons pas dans le détail ici.

2. Avec ce type d'approche, dans le cadre d'une analyse scientifique, il est crucial de déterminer si la différence observée est significative ou si elle peut être attribuée au hasard (deux échantillons ne contiendront jamais exactement le même nombre d'occurrences, ce qui ne signifie pas nécessairement qu'il y ait là une différence systémique qui doive être prise en compte). Pour chaque comparaison effectuée, les données chiffrées obtenues ont donc été validées statistiquement. Dans la mesure où nous comparons deux échantillons indépendants, nous avons ici eu recours au test de significativité entre deux proportions indépendantes (calcul d'un *z-ratio* et d'une double *p-value* — seuil de significativité retenu p = 0.01. Les analyses statistiques ont été effectuées par le biais su site internet VassarStats, faculty.vassar.edu/lowry/). Cette validation statistique ne concerne évidemment pas le traducteur professionnel (voir section suivante).

3. Naturellement, il importe de bien délimiter le *tertium comparationis* : il s'agit de comparer la fréquence des structures en *there/il y a* permettant d'introduire un élément nouveau en discours. Sont donc exclues les structures présentationnelles en *there* (avec un autre verbe que *be*) et l'emploi dit « prépositionnel » de *il y a* (p. ex. *il y a deux heures*). Voir Cappelle et Loock (2013) pour un exposé détaillé de la méthodologie, notamment sur la façon d'éliminer les exemples dits « bruyants » comme *Il n'y a jamais mis les pieds* ou *There is your book*.

4. Ainsi, pour le Pape Jean-Paul II, on constate que le nombre de possibilités est plus important : *Pope John Paul II*, *Pope John Paul*, *John Paul II*, *The Pope*, *Karol (Jozef) Wojtyla*, *Karol*, *Wojtyla*.

RÉSUMÉS

Dans cet article, nous expliquons la façon dont les corpus électroniques peuvent être utilisés comme outils d'aide à la traduction (outils de TAO), aux côtés des autres outils plus

« traditionnels » comme les dictionnaires en ligne, les glossaires électroniques, les mémoires de traduction ou encore la traduction automatique. Spécifiquement, il s'agit de déterminer à quelle (s) étape(s) du projet de traduction (recherche documentaire, traduction, post-édition/révision) le recours aux différents types de corpus peut être le plus utile aux traducteurs, mais aussi de formuler des propositions afin de résoudre quelques problèmes d'ordre ergonomique posés par ces banques de données. Nous distinguons pour cela différents types de corpus (les corpus dits « de référence » comme le *British National Corpus* ou le *Corpus of Contemporary American English* ; les corpus de langue cible originale que l'on peut compiler soi-même pour la circonstance — *do-it-yourself* ou *DIY corpora* ; les corpus comparables et parallèles) et montrons, exemples concrets à l'appui, qu'ils apportent différentes informations d'ordre linguistique au traducteur. En particulier, nous nous attardons sur l'utilisation de deux types de corpus encore peu répandue à l'heure actuelle : le corpus DIY monolingue en langue cible originale, qui permet d'augmenter la fluidité de la langue cible, et le corpus DIY comparable de langues originales (langue source originale – langue cible originale), qui permet la prise en compte de l'usage (lexical, phraséologique, discursif) des textes très spécialisés (« minilectes »).

In this article, we explain how electronic corpora can be used as translation tools (CAT tools) to complement other, more "traditional" tools like on-line dictionaries, electronic glossaries, translation memories, or machine translation. Specifically, this article aims to determine the different steps of a translation project (documentary research, translation, post-edition/revision) for which using different kinds of corpora can be the most useful for translators, but also aims to make suggestions to try and remedy some of the ergonomics-related issues associated with the use of such linguistic databases. In order to do so, we distinguish between different types of corpora (reference corpora like the *British National Corpus* or the *Corpus Of Contemporary American English*; self-collected corpora of original target language texts (*do-it-yourself* or *DIY corpora*); comparable and parallel corpora) and show, through concrete examples, that they provide translators with different kinds of linguistic information. In particular, we focus on the use of two types of corpora which is still quite rare today: monolingual DIY corpora in the original target language, which allow for the increase of the target texts' naturalness, and comparable DIY corpora of original languages (original source language – original target language), which help to reveal usage (lexical, phraseological, discursive) of highly specialized texts ("minilects").

INDEX

Mots-clés : corpus, outil d'aide à la traduction, fluidité, qualité, usage, minilecte
Keywords : corpus, translation tool, fluidity, quality, usage, minilect

AUTEUR

RUDY LOOCK

Université Lille Nord de France, UMR STL 8163 du CNRS