Bases de données et traductions : Aspects quantitatifs et qualitatifs

Yvon Keromnes

Bases de Données et Traductions : Aspects Quantitatifs et Qualitatifs

Yvon Keromnes
Université de Metz & Atilf-CNRS Nancy

1 Introduction

Nous nous proposons dans cet article d'aborder d'un point de vue méthodologique l'utilisation de bases de données en linguistique et en théorie de la traduction. En matière d'études de corpus, la tendance actuelle est à l'exploitation extensive de corpus étendus, ce qui correspond à une logique à la fois pratique et théorique ; pratique tout d'abord, puisque la tendance reflète l'évolution des logiciels et matériels informatiques permettant un traitement systématique et toujours plus rapide de bases de données toujours plus importantes ; théorique aussi, parce que cette évolution recouvre des enjeux épistémologiques liés à la taille et à la variété des corpus, ainsi qu'aux instruments qui leur sont appliqués : les enseignements que l'on pourra tirer d'un corpus sont souvent liés à une taille critique de celui-ci ; ainsi peut-on lire dans la préface du *Collins Cobuild Dictionary*, que son équipe de lexicographes, qui dans les années 80 s'enorgueillissait de son corpus de 20 millions de mots, avait dû déchanter en constatant l'impossibilité de proposer une description adéquate de la totalité de la langue anglaise à partir d'un tel corpus. Un corpus de 200 millions de mots permettrait d'y parvenir.

Dès lors, la taille semble être le paramètre déterminant pour la recevabilité des corpus. Dans le domaine français, *Frantext*, qui atteint également 200 millions de mots dans sa version intégrale, constitue la ressource privilégiée pour les textes littéraires. Pour la presse, la référence est le journal *Le Monde*. Or, l'utilisation de ces deux ressources n'est pas sans poser des problèmes, en premier lieu ceux de l'uniformité et de la représentativité : dans quelle mesure les textes les plus classiques reflètent-ils la langue française ou même simplement la littérature ? Et si la langue journalistique possède ses spécificités, *le Monde* est-il prototypique à cet égard ? Enfin, que peuvent nous apprendre des statistiques sur ce type de corpus, outre la confirmation de leur homogénéité ?

Pour certains, la solution à ces problèmes sera dans la course en avant, l'application d'outils mathématiques toujours plus sophistiqués à des corpus toujours plus grands. Dans cette perspective, seul le web lui-même serait bientôt un corpus digne d'intérêt... à condition qu'il poursuive une croissance accélérée.

2. Corpus et Traduction

Quelle est la pertinence de ces problèmes... et de leurs solutions pour des études de corpus ayant pour but de théoriser la traduction ? G. Bourquin (1984) a pu écrire à ce sujet :

> « Pour appréhender clairement les variables interlinguistiques pertinentes pour la traduction, une étude sur un corpus très étendu est nécessaire ».

Pourtant, sans vouloir rejeter la notion de validité liée à la taille du corpus, c'est une approche très différente que nous voulons proposer ici, avec une base de données élaborée dans le cadre de notre thèse (Keromnes, 2000) consacrée au fonctionnement des formes verbales dans la narration en allemand, en en anglais et en français, et qui a la particularité de permettre la comparaison entre un texte original allemand (*La Métamorphose*, de F. Kafka) et huit traductions publiées de cette nouvelle, quatre traductions en anglais et quatre en français, soit au total, si l'on associe l'original à l'ensemble paraphrastique, neuf versions d'un même texte. La conception de cette base de données de taille très modeste (environ 190.000 mots pour le corpus non balisé) découle des principes et objectifs exposés ci-dessous.

Le premier principe est notre approche paraphrastique de la traduction. L'opération de traduction articule quatre dimensions, celles de la langue source, du texte source, de la langue cible et du texte cible, ce dernier étant notablement instable[1] ; c'est en faisant reposer essentiellement la traduction sur la seule mise en rapport de deux systèmes linguistiques que Vinay & Darbelnet (1958 : 24) pouvaient écrire :

> « Il est permis de supposer que si nous connaissions mieux les méthodes qui gouvernent le passage d'une langue à l'autre, nous arriverions dans un nombre toujours plus grand de cas à des solutions uniques ».

Aujourd'hui, au contraire, nous dirons qu'à partir d'un texte original, la pluralité des traductions ne doit pas être considérée comme une exception, mais comme la règle. Et c'est cette pluralité qui constitue notre objet d'étude. Nous rejoignons sur ce point M. Ballard (1995 : 11) :

> « Une étude traductologique se doit, à nos yeux, de tenir compte des manières de traduire, qui sont diverses et parfois en désaccord ».

L'articulation des quatre dimensions de la traduction ne peut selon nous être véritablement appréhendée dans une problématique de la paraphrase (cf. C. Fuchs, 1975), non seulement dans une perspective interlinguistique, mais également intralinguistique ; le fait que l'on ne puisse aujourd'hui prétendre, de façon générale, à l'existence d'une solution unique en matière de traduction impose de prendre en compte les variantes éventuellement contradictoires, mais le texte source doit lui-même être mis en regard des différentes variantes qui peuvent lui être opposées[2]. Au centre des ensembles paraphrastiques intralinguistique ou extralinguistique se trouve le texte source, en regard duquel chaque paraphrase est produite, ce qui a pour effet de limiter fortement le nombre de paraphrases possibles[3]. D'où l'idée d'une base de données permettant de comparer systématiquement plusieurs traductions, en l'occurrence dans deux langues cible, d'un même texte. Et notre conception de la paraphrase nous donne à penser que la comparaison des ensembles paraphrastiques anglais-français sera aussi utile que les comparaisons allemand-anglais et allemand-français, l'original allemand constituant dans le premier cas un tertium comparationis implicite. Autrement dit, dans cette base de données, le facteur d'extensivité, en tant que critère de validité des études de corpus, cède la place à un double facteur de

complexité : au texte source correspondent quatre textes cible dans deux langues cible, eux-mêmes objets de comparaison.

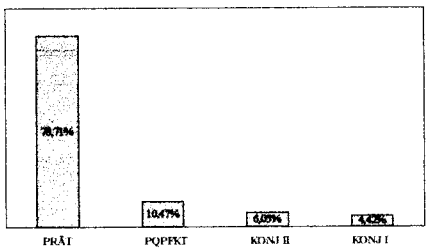
Notre objet d'étude étant les formes verbales dans la narration et la traduction, la base de données permettant cette étude doit répondre à deux objectifs antinomiques, permettre d'observer la variabilité des représentations sémantiques à partir d'un référent supposé stable, et d'autre part conserver la structure de surface, dans laquelle plusieurs facteurs touchant à la textualité sont en jeu (ordre des termes, cohésion, cohérence...). D'où notre décision de retenir finalement le format du tableur Excel pour cette base de donnée. Texte original et traductions y sont découpés en *équivalents propositionnels* (ci-après EP) selon une conception de la *proposition* exposée dans François & Keromnes (1995 : 45-46), et qui se situe à mi-chemin entre celle de Kintsch & Van Dijk (1975 : 99) et celle de la grammaire traditionnelle : nous proposons un découpage propositionnel reposant sur un élargissement de l'acception traditionnelle de 'proposition' à tous les syntagmes (nominaux, adjectivaux, adverbiaux, participiaux) à fonction prédicative comportant un complément ou une détermination adverbiale et/ou étant introduits par une préposition non régie : [A person [knowing a lot]P] vs [A knowing person] : 'knowing a lot' étant reconnu comme syntagme dépendant, le premier exemple est compté comme EP alors que le second ne l'est pas. De la sorte, nous sommes à même d'observer la traduction de formes verbales finies par d'autres formes verbales finies que la traduction de formes verbales non finies, voire de formes non verbales, par des formes verbales finies, et inversement.

Enfin, d'un point de vue méthodologique, nous pensons qu'une base de données telle que nous l'envisageons doit répondre à un *critère de complétude* en présentant l'intégralité d'un texte et des traductions dont nous disposons. En effet, d'une part, pour le traducteur, le texte à traduire constitue en soi une unité, et on ne peut prétendre véritablement cerner le fonctionnement des formes verbales entre un texte et ses traductions si l'on se limite à étudier des fragments : comme l'a montré Weinrich (1973), certaines formes verbales sont plus représentées à certaines étapes d'un récit, en français par exemple, des imparfaits seront plus nombreux en début ou en fin de récit, des passés simples dans le déroulement[4]. Par ailleurs, en présentant les traductions dans leur intégralité, nous nous interdisons de préjuger de la recevabilité des observations que nous pourrons faire sur elles : pour juger d'une tendance propre à un système linguistique particulier comparé à un autre, ou pour juger de la pertinence d'une traduction donnée d'une forme verbale, il faut disposer des différents choix effectivement faits en discours. Sans quoi la tentation serait trop grande de ne retenir que les tendances qui correspondent à ce qu'on veut démontrer.

3. Extractions et dénombrements

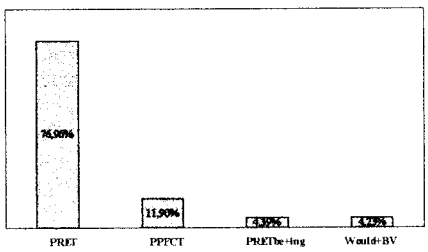
Comme pour les corpus de taille plus importante, l'exploitation de celui-ci revêt en premier lieu une dimension quantitative. Dans Excel, chaque texte, placé dans une colonne, conserve sa lisibilité, tandis que chaque ligne fait correspondre à un EP de l'original allemand quatre EP en français ou en anglais. Les EP sont pourvus de différentes étiquettes syntaxiques et sémantiques, dont la principale, l'indication de la marque de temps/mode/aspect du morphème verbal, mais aussi les numéros de paragraphes, phrases et proposition, le niveau de subordination, catégorie et fonction syntaxiques des EP...). La base de données ainsi

constituée intègre donc à la fois le corpus et les instruments de son exploitation, ce qui facilite la reproductibilité des observations : en mode filtre, chaque tête de colonne présente un menu déroulant affichant le contenu de la colonne ; par la sélection d'une ou plusieurs valeurs de ce menu, nous pouvons en extraire instantanément toutes les occurrences de ces valeurs dans la colonne. Ainsi, par la sélection successive des différents marqueurs de temps, mode ou aspect dans la colonne appropriée (TMA) correspondant au texte original, nous pouvons représenter la répartition des formes verbales dans ce texte.

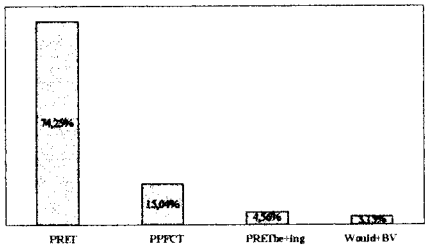


Graphique 1 : répartition des formes verbales finies principales[5] dans *die Verwandlung* (Präteritum, Plusquamperfekt, Konjunktiv II, Konjunktiv I)

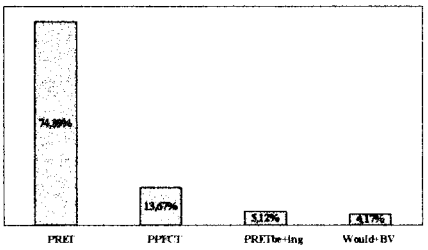
De même pour chacune des traductions (ici les traductions anglaises) :



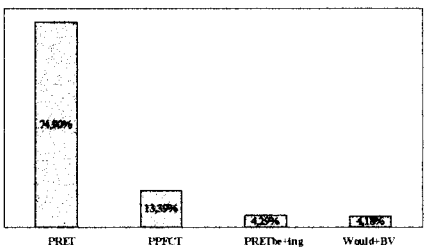
Graphique 2 : traduction de Muir & Muir



Graphique 3 : traduction de Underwood



Graphique 4 : traduction de Pasley



Graphique 5 : traduction de Corngold

Et alors que l'on peut s'attendre à ce que chaque traducteur tende à se distinguer de ses prédécesseurs, l'utilisation des morphèmes verbaux dans les traductions se montre relativement homogène. Si nous appliquons à présent le même filtre TMA dans les quatre traductions, nous obtenons la répartition des formes verbales faisant l'objet de décisions unanimes. Ce taux d'accord moyen nous donne une nouvelle indication sur l'emploi de ces formes verbales :

Formes verbales finies	Taux d'accord moyen
PRET	77,70%
PPFCT	58,57%
PRETbe+-ing	34,54%
Would+BV	37,56%

Tableau 1 : taux d'accord dans le choix des formes verbales en anglais

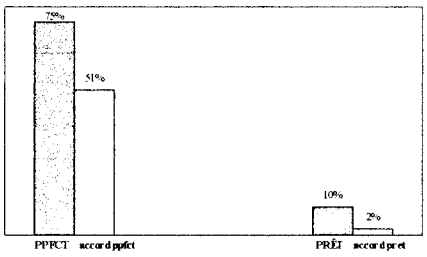
Nous voyons que si le choix d'un *prétérit* est le plus souvent unanime, celui d'un *pluperfect*, deuxième forme verbale la plus représentée, offre matière à une variation bien plus grande, et que les quatre traductions ne se retrouvent qu'à peu près une fois sur trois pour l'emploi des deux autres formes verbales. Ce taux d'accord est révélateur de l'existence de contraintes linguistiques plus ou moins fortes pesant sur le choix des traducteurs[6].

Si nous appliquons un filtre TMA dans l'original pour extraire, par exemple, les occurrences du *Plusquamperfekt*, deuxième forme verbale la plus représentée, puis un deuxième filtre aux différents marqueurs TMA d'une traduction, nous obtenons la répartition des traductions du *Plusquamperfekt* dans ce texte cible, aussi bien dans leurs aspects quantitatifs (nombre d'occurrrences de *pluperfect simple* ou *périphrastique*, de *prétérit* etc. pour une traduction en anglais) que qualitatifs (nombre de possibilités de traduire différemment un prétérit allemand) ; et là encore, nous pouvons comparer les traductions entre elles.

	Muir & Muir	Underwood	Pasley	Corngold
PPFCT	143	163	171	173
PRET	50	19	23	22
Would+BV	1	0	0	0
Would+Have+ -en	2	2	2	1
PPFCTbe+-ing	4	2	3	5
PRETbe+-ing	0	0	2	1

Tableau 2 : traductions du *Plusquamperfekt* (212 occurrences) en anglais

Nous n'indiquons ici que les traductions par des formes verbales finies (il manque donc les traductions par des formes verbales non finies, par des formes non verbales ou encore les non-traductions) mais nous voyons déjà d'une part la diversité qualitative, et d'autre part la concentration quantitative des traductions sur deux formes verbales. Là encore, si nous comparons les moyennes d'occurrence de ces deux formes verbales à leur taux moyen d'accord, nous obtenons une représentation différenciée de cet emploi :



Graphique 6 : traduction du *Plusquamperfekt* en anglais, emplois comparés des *pluperfect* et *prétérit* et taux d'accord moyens

Ce rapprochement de la répartition des traductions du *Plusquamperfekt* et des taux d'accord dans les traductions anglaises fait apparaître une forte contrainte pour les traduction par des *pluperfects* alors que l'emploi des *prétérits*, qui correspond tout de même à 10% des traductions du *Plusquamperfekt*, fait partie des choix non contraints.

L'utilisation de filtres multiples permet enfin de comparer la variabilité des traductions selon des critères syntaxiques (nature des propositions, degré d'enchâssement...) et textuels (début ou fin de paragraphe...).

D'un point de vue linguistique, une étude quantitative de la répartition des formes verbales permet de mieux jauger des proportions de l'identique et du différent ; nous l'avons vu dans le cas de l'anglais, mais entre l'anglais et le français, on constatera également qu'un *prétérit simple* pourra correspondre non seulement à un *imparfait* ou à un *passé simple*, mais aussi à un *plus-que-parfait*, un *conditionnel*, un *subjonctif*, et même un *présent*, un *infinitif* ou un *participe présent* ; en revanche, dans 90% des cas, en face du *prétérit simple*, ce sera bien un *imparfait* ou un *passé simple* que nous trouverons.

4. Conclusion

Nous avons voulu montrer, à l'encontre d'une tendance actuelle dans les études de corpus, l'intérêt linguistique et traductologique d'une base de données pour laquelle le critère de validité instrumentale ne réside pas dans la taille, mais dans la possibilité d'observer, sur des ensembles paraphrastiques et selon de multiples paramètres, la variabilité, tant qualitative que quantitative, des traductions, ainsi que la possibilité d'extraire à chaque fois les occurrences correspondant à des emplois linguistiquement contraints ou au contraire stylistiquement libres, deux possibilités associées à ce que nous avons appelé *critère de complétude*. Les limites de la démarche sont claires : il s'agit d'un texte littéraire particulier, dans lequel nos observations ne portent que sur le discours narratif, et les traductions observées sont uniquement orientées de l'allemand vers l'anglais et vers le français (c'est-à-dire vers une explicitation du marquage aspectuo-temporel). Pourtant, nous pensons que la description exhaustive des 2380 équivalents propositionnels de l'original et de leurs traductions, la multiplication des perspectives que crée la mise en correspondance de 9 variantes d'un même texte sur trois langues compense largement ces limites, pour faire de cette base un instrument tout à fait pertinent de validation... ou de falsification théorique.

Références

1) Corpus

Texte original :
Kafka F. 1983 (1ère publication : 1915). Die Verwandlung, in : Das Urteil, Fischer, Hamburg.
Traductions anglaises :
Corngold S. 1996 (copyright traduction : 1972). The metamorphosis, a Norton critical edition, New York, London.

Muir W. & Muir E. 1995 (copyright traduction : 1949). Kafka, the complete short stories. London : Minerva.
Pasley M. 1992. The Transformation and Other Stories. London : Penguin.
Underwood J. A. 1991 (copyright traduction 1981). Franz Kafka Stories 1904-1924. London : Cardinal.
Traductions françaises :
David C. 1990 (copyright traduction 1989). La métamorphose et autres récits. Paris : Gallimard.
Lortholary B. 1988. La métamorphose, Description d'un combat. Paris : GF-Flammarion.
Vergne-Cain B. & Rudent G. 1988. La métamorphose. Paris : le Livre de Poche.
Vialatte A. 1955. La métamorphose. Paris : Gallimard.

2) Linguistique et traductologie

Ballard M. 1987. La traduction : de l'anglais au français. Paris : Nathan.
Ballard M. (ed). 1995. Relations discursives et traduction. Lille : Presses universitaires de Lille.
Bourquin G. 1984. Semantic Relations, Modality, Time-Tense. A resolution study on French, Document C.E.E., Eurotra Report ETL-F-1 (4), Luxembourg, 1984, 220 pp.
Demanuelli C. & Demanuelli J. 1991. Lire et traduire : anglais-français. Paris : Masson.
François J. & Keromnes Y. 1995. De la fidélité en traduction littéraire : esquisse d'une méthode d'évaluation comparative, in : Lectures, Hommage à Geneviève Hily-Mane, Presses Universitaires de Reims, pp. 41-68.
Fuchs C. 1994. Paraphrase et énonciation, Paris : Ophrys.
Keromnes Y. 2000. Formes verbales, narration et traduction : étude d'une nouvelle de F. Kafka et de quatre traductions de cette nouvelle en anglais et en français, Thèse de doctorat, Université Nancy 2.
Keromnes Y. 2005. Equivalences paraphrastiques : une question de perspective, in : *Recherches en Linguistique et Psychologie cognitive* 21, Presses Universitaires de Reims. pp. 241-261.
Kintsch W. & van Dijk T. 1975. Comment on se rappelle et on résume des histoires, in : *Langages* 40, Paris : Larousse. pp. 98-116.
Mel'cuk I. 1993. Cours de morphologie générale, vol. 1, Presses de l'Université de Montréal, CNRS Editions.
Weinrich H. 1973. Le Temps (trad. française M. Lacoste), Paris : le Seuil.

Notes

[1] Cf. C. Demanuelli & J. Demanuelli (1991 : 7) ; également M. Ballard (1987 : 45), qui explique par l'instabilité du texte à produire la variabilité de l'unité de traduction.

[2] Pour plus de détails sur cette approche paraphrastique de la traduction, cf. Keromnes (2000, 2005).

[3] Nous sommes donc ici à l'opposé de l'approche des paraphrases en termes de synonymie relative de I. Mel'cuk (1993 : 44), qui dans son modèle de génération paraphrastique, parvient en théorie à un ensemble de 50 millions d'éléments à partir d'une simple phrase extraite d'une dépêche d'agence de presse.

[4] L'importance de cette prise en compte de la dimension textuelle ne se limite d'ailleurs pas à l'étude des formes verbales. Une étude du connecteur *and* que nous avons mené sur ce même corpus fait apparaître des emplois tout à fait particuliers en fin de paragraphe, de section et de la nouvelle elle-même.

[5] Les formes verbales représentant moins de 1% de l'ensemble des formes verbales finies (par ex. le présent, 6 occurrences, 0,3% de l'ensemble) ne sont pas représentés sur ce graphique.

[6] On pourrait penser que le taux d'accord est d'autant plus fort que la forme verbale est fortement représentée, or il n'en est rien. Ainsi, dans les traductions françaises, alors que le *passé simple* est sensiblement moins employé que *l'imparfait*, le taux d'accord moyen est de 72,50% pour le premier et de 62,85% pour le second : les contraintes aspectuelles liées à la progression narrative sont donc bien plus fortes pour le *passé simple* que pour l'*imparfait*.