# Corpus parallèles massifs et traduction. Quelles pépites se cachent dans les séquences-pivots des moteurs de traduction phrase-based ?

Antonio Balvet

# CORPUS PARALLÈLES MASSIFS ET TRADUCTION : QUELLES PÉPITES SE CACHENT DANS LES SÉQUENCES-PIVOTS DES MOTEURS DE TRADUCTION *PHRASE-BASED ?*



## INTRODUCTION

Le domaine de la Traduction Automatique (TA) a subi des bouleversements majeurs au cours des vingt dernières années, avec l'avènement de la traduction guidée par les exemples observés en corpus, qui a marqué un tournant empirique dans la mécanisation de la traduction. Une première étape avait été franchie à la fin des années 1990, avec les premiers systèmes de traduction statistique (SMT). Ce mouvement général semble devoir s'amplifier, avec l'avènement de moteurs de TA dits neuronaux (NMT), à même de traduire en un temps record des ouvrages scientifiques avec un taux d'intervention manuelle de post-édition réduit[1]. La disponibilité de volumes jusque-

[1] La start-up Quantmetry, en collaboration avec la société DeepL, commercialisera courant 2018 une version traduite en français de l'ouvrage de référence sur l'apprentissage « profond », Goodfellow *et al*. (2016), soit plus de 700 pages traduites en seulement 12 heures. D'après Quantmetry, le recours à une traduction neuronale de bonne qualité a rendu possible une mise sur le marché en quelques mois, contre plus d'un an dans le cas d'une traduction manuelle.

là inenvisageables de corpus de traductions, ou corpus parallèles, structurés et alignés de façon standardisée et inter-opérable, a permis l'émergence de stratégies de traduction guidées par l'usage et non plus par des connaissances explicites (règles) fournies par des experts de la traduction.

Le paradigme des approches par règles, dominant jusqu'à la fin de années 1990, semble ainsi avoir cédé la place à une autre famille d'approches, guidées par les données et reposant sur des modèles probabilistes de correspondance langue-source → langue-cible. En effet, les modèles statistiques de traduction ont remplacé la coûteuse phase d'édition manuelle de ressources lexicales et de règles d'analyse syntaxique locales par l'induction automatique de modèles de traduction à partir de milliers d'exemples de traductions. Les systèmes les plus récents de cette famille, reposant sur des réseaux de neurones artificiels, cherchent à améliorer les performances des systèmes classiques de SMT, notamment en constituant des représentations affinées des différents déclencheurs d'emploi d'une forme linguistique donnée (mot, groupe de mots) en fonction du contexte. Par ailleurs, ils visent à contourner le verrou que représente l'explosion combinatoire liée au développement de moteurs de traduction par paires de langues, via la « zero-shot translation », ou traduction transitive, présentée dans Johnson *et al.* (2016), voire à s'affranchir des corpus parallèles (Lample *et al.*, 2018).

Dans le présent article nous nous pencherons sur les moteurs de traduction par séquences-pivots, ou *Phrase-Based Machine Translation* (PBMT). À première vue, il est tentant de ne voir dans la PBMT qu'une simple variante de la SMT, en ce qu'elle cherche à constituer des règles de transfert au niveau de séquences de mots (les *phrases*) plutôt qu'à celui des mots isolés. Ce faisant, la PBMT pose toutefois une question fondamentale, qui dépasse les enjeux des gains de performance. En effet, ces moteurs de TA ont pour objectif, étant donné un corpus parallèle, d'en extraire des équivalents traductionnels qui couvrent des séquences de plusieurs mots, en n'ayant recours ni à une analyse syntaxique, ni à une interprétation sémantique (encore moins pragmatique, stylistique, ou autre). Nous posons que les améliorations apportées à la qualité des traductions par séquences-pivots, par rapport aux moteurs SMT conventionnels, reflètent des propriétés de structure des langues naturelles : si l'opération d'extraction de *phrases* est possible entre deux langues, c'est bien qu'elles partagent, au moins en partie, des éléments de structure (les mêmes mots dans les mêmes configurations). Ceci implique, en outre, que ces séquences constituées comme pivots soient suffisamment fréquentes et régulières pour qu'elles émergent mécaniquement d'un processus d'appariement entre langues source et cible. En d'autres termes, nous avançons que le succès des moteurs de TA guidés par les données (SMT, PBMT, NMT, ainsi que traduction par analogie), au-delà de l'amélioration de la qualité traductionnelle, interroge le

traducteur/linguiste sur le statut des représentations linguistiques humaines, les effets de fréquence, et partant le statut des outils statistiques pour l'analyse des langues naturelles.

Dans un premier temps, nous présenterons le domaine général des moteurs de TA guidés par les données, par opposition aux moteurs guidés par des connaissances linguistiques explicites. Nous aborderons ainsi l'architecture générale des moteurs de SMT, puis nous nous attacherons aux spécificités des moteurs de PBMT. Par la suite, nous examinerons en détail des exemples de phrases, ou séquences-pivots, induites à partir de corpus parallèles. Cette section sera l'occasion de mesurer à la fois la pertinence de l'approche pour le problème de la TA, mais également d'identifier des manques parfois rédhibitoires, qui tiennent justement à une cécité à la structure linguistique au-delà des apparences immédiates. La dernière section s'ouvrira sur une discussion des unités manipulées par la PBMT et sur l'intérêt de ces unités pour le traducteur/linguiste, au-delà de leur application directe.

## SMT, PBMT, EBMT, NMT : LE TRIOMPHE DES APPROCHES EMPIRIQUES

Dans cette section, nous examinons les fondements techniques et conceptuels de différentes approches de la TA : SMT, PBMT et NMT, mais également traduction par analogie,

habituellement classée dans la famille des moteurs de TA à base d'exemples (EBMT). En premier lieu, signalons que Wilks (2008) établit une distinction entre EBMT et SMT pour des raisons à la fois historiques et technologiques : les premiers systèmes EBMT exploitaient des relations analogiques entre mots et contextes, à l'aide de règles formelles (raisonnement analogique). De ce point de vue, les premiers systèmes de SMT, tels que présentés par les équipes d'IBM dans Brown *et al.* (1988), ne s'inscrivaient pas dans une stratégie de traduction par analogie (*Translation by Analogy*) telle que problématisée par Nagao (1984). Toutefois, bien que la stratégie de la traduction par analogie n'intègre pas nécessairement une dimension statistique, elle repose néanmoins sur des corpus de traduction. Nous proposons par conséquent une catégorie générale, comprenant les systèmes guidés par les connaissances explicites (règles), par opposition aux approches empiriques (systèmes guidés par les données). Par ailleurs, nous rejoignons Wilks (2008) quand il voit dans le projet de la SMT un changement majeur de paradigme tant technique que théorique.

> *The shock of the "IBM statistical MT movement" [...] seems to have passed, and the results are indecisive as regards a fully statistical MT project. [...] What was given substantial new life by the IBM project was a much wider empirical, data-driven or statistical linguistics: a methodology in decline since Chomsky's original attacks on it, and which came back into MT work through its successful application to speech recognition,*

> *by the same IBM team, among others (Wilks, 2008, p. 4).*

Au-delà du domaine de la TA, les approches guidées par l'usage semblent convaincre de plus en plus de linguistes, y compris théoriques. Or, jusqu'à présent, ces approches fondées sur l'usage effectif se heurtaient à la position farouchement anti-probabiliste des linguistes formalistes tels que N. Chomsky :

> *I think we are forced to conclude that [...] probabilistic models give no particular insight into some of the basic problems of syntactic structure (Chomsky, 1957, p. 17).*

Dit d'une autre façon :

> *It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term (Chomsky, 1969, p. 57).*

À la lecture de ces citations, on peut comprendre pourquoi Wilks parle de la chute d'un « rideau de fer intellectuel » (*the fall of an intellectual Berlin Wall*) au sujet de l'adoption du tout-statistique par les équipes d'IBM. Le recours aux approches statistiques avait, dans un premier temps, été validé dans le domaine de la reconnaissance vocale. Il a, par la suite, percolé au-delà de ce domaine d'application, pour s'étendre à la TA, mais également au-delà (par exemple : reconnaissance de formes, de caractères).

# PRINCIPES FONDAMENTAUX DES MOTEURS DE SMT

Bien que les premières implémentations viables de moteur de SMT datent de la fin des années 1980, les fondements conceptuels de la traduction statistique sont déjà posés dans King (1961) : il défend l'idée que les langues naturelles sont fondamentalement redondantes, il est donc, au moins en principe[2], possible de tirer parti de cette redondance en cherchant à deviner le mot suivant le plus probable par rapport au mot courant, dans un texte. L'article de King est passé relativement inaperçu à l'époque de sa parution, dans un contexte où dominaient alors les approches guidées par les connaissances, tant en IA qu'en TAL : systèmes-experts, analyseurs syntaxiques et générateurs automatiques de phrases à partir de formalismes syntaxiques inspirés de la Grammaire Générative[3]. L'absence de corpus de traductions au format électronique, disponibles en volumes suffisants, rendait par ailleurs difficile toute expérimentation, même limitée, d'une approche probabiliste de la TA.



[2] À l'époque où l'article de King paraît, l'induction de modèles statistiques de probabilités de transitions entre mots reste une vue de l'esprit : aucun corpus de taille suffisante, exploitable par une machine n'est encore disponible.

[3] Le schisme entre Traitement Automatique des Langues et Grammaire Générative n'aura lieu que bien plus tard, à la fin des années 1970.

Il faut attendre plus de 20 ans entre l'article de King et celui de Brown *et al.* (1988), qui expose les principes fondamentaux des moteurs de SMT. L'équipe d'IBM à l'origine du premier moteur de SMT, CANDIDE, y expose comment l'application au problème de la TA de méthodes issues de la reconnaissance de la parole est non seulement viable, mais est également susceptible d'apporter des gains de performances, par rapport à un système à base de règles. Dans le contexte de l'époque, dominé par les approches guidées par les connaissances, et face à la suprématie commerciale du moteur de Systran, la proposition de Brown et ses collègues semble révolutionnaire. En effet, la stratégie exposée permet, à condition de disposer de corpus parallèles en quantité suffisante – c'est-à-dire d'un échantillon représentatif du problème – d'induire des modèles de traduction pour n'importe quelle paire de langue, en appliquant les mêmes principes fondamentaux, quelles que soient les langues. Cette stratégie met au centre du processus de traduction les corpus parallèles, tels que les corpus bilingues Hansard du Parlement du Canada. L'accent est également mis sur le volume : les premières expérimentations de l'équipe d'IBM portent sur un corpus de 30 millions de mots, à une époque où les premiers moteurs de SMT présupposent encore un équipement informatique conséquent (serveurs de grande capacité, temps de calcul importants).

Dans un monde alors dominé, tant en TA qu'en IA, par les approches guidées par les connaissances, la publication de Brown *et al.*

(1988) jette un véritable pavé dans la mare. En effet, l'équipe d'IBM fait le pari d'une approche linguistiquement naïve, faisant l'impasse sur tout « mécanisme intermédiaire (langue) qui encoderait le "sens" du texte source ». Cependant, la condition essentielle pour que la stratégie proposée fournisse des traductions exploitables est la disponibilité d'un corpus de traduction. L'autre prérequis technique concerne bien évidemment l'outillage informatique à même d'induire les modèles statistiques requis : essentiellement, modèle des probabilités de transition entre mots de la langue-source, et modèle de la distorsion subie au cours de la traduction[4]. Comme le résument les auteurs :

> *Our procedure will be based on a model (an admittedly crude one) of how English words are generated from their French counterparts.*
>
> « Notre procédure est fondée sur un modèle (simpliste) prédisant comment les mots anglais sont générés à partir des mots français ».

L'objectif premier de la stratégie présentée dans Brown *et al.* (1988) est donc l'induction, par observation d'exemples de traductions, de la probabilité conditionnelle de traduire un mot français par un mot anglais donné. L'algorithme présenté est volontairement « naïf » dans le sens où il ne présuppose aucune connaissance linguistique explicite. Par ailleurs, l'approche

[4] Un modèle des probabilités de transitions entre mots pour la langue-cible sera également utilisé, plus tard, pour améliorer la forme de surface des phrases ainsi traduites.

présuppose qu'à chaque phrase française correspond une phrase anglaise. De façon plus dérangeante pour un linguiste/traducteur, cette stratégie présuppose également que chaque mot français ait un équivalent traductionnel en anglais. Les auteurs sont conscients de la simplification outrancière que constituent ces choix : en pratique, il est quasiment impossible d'aboutir à une correspondance 1-1 entre mots français et anglais[5]. On le voit, l'approche proposée repose entièrement sur le paradigme de la traduction vue comme le décodage d'un message via un canal bruité (cf. infra), ainsi que sur une approche volontairement simpliste du problème de la traduction. Bien que des ajustements soient nécessaires, notamment en termes statistiques (ex. : lissage des fréquences), l'essentiel de la méthode repose sur des éléments relativement faciles à extraire d'un corpus parallèle. La force de la proposition de Brown *et al.* réside, justement, dans l'intégration de méthodes statistiques ayant fait leurs preuves dans d'autres domaines (en l'occurrence la reconnaissance vocale), motivées de façon générale par la théorie de l'information.

## LE MODÈLE DU CANAL BRUITÉ

Les moteurs de TA guidés par les données, et en particulier les moteurs SMT, font appel au modèle du canal bruité proposé par Shannon (1948) et étendu à la communication en général

[5] Si traduire devait se borner à apprendre les correspondances mot-à-mot entre langue-source et langue-cible, la traduction ne serait effectivement que du transcodage.

(communication humaine, télécommunications), et à la traduction en particulier, dans Weaver (1949). Dans ce modèle, un message est transmis d'un émetteur vers un récepteur. Au cours de sa transmission, il se trouve bruité (crypté, ou tout simplement en langue-source), et la tâche du traducteur, d'après Weaver, est de trouver la clé qui permettra le décodage (traduction en langue-cible). En traduction par moteur statistique, on cherche à trouver un modèle de la distorsion subie par le message, autrement dit un modèle de traduction pour une paire de langues donnée. On dispose également d'un modèle de la forme probable du message, autrement dit un modèle de langue (par exemple : des transitions de probabilités entre mots de la langue-source). Le modèle de traduction peut prendre plusieurs formes :

- une variante d'un lexique de traduction, induite automatiquement à partir de grands volumes d'exemples de traduction, pour les systèmes SMT reposant sur un alignement mot à mot ;
- un lexique de traduction composé, cette fois, de séquences de mots et non plus de mots isolés, pour les systèmes de PBMT ;
- un modèle de transfert codé dans les pondérations des cellules composant les couches cachées d'un moteur de NMT.

Dans le cas du lexique de traduction créé manuellement, le modèle de traduction est généralement limité au mot, associé spontanément par le traducteur/linguiste à un ensemble de

contextes d'emploi. Pour les modèles acquis automatiquement à partir de corpus, une métrique d'adéquation permet d'explorer toutes les possibilités de traduction, pour ne retenir que celles qui dépassent un seuil déterminé de manière empirique. Le corpus de traductions permet, dans cette configuration, de paramétrer de façon automatique le modèle de traduction.

### Alignement de segments phrastiques

À supposer que des corpus parallèles de taille suffisante soient disponibles, encore faut-il aligner les phrases de la langue-source avec celles de la langue-cible. Or, l'alignement manuel des phrases d'un corpus de plusieurs dizaines de millions de mots est une tâche extrêmement longue, sujette aux erreurs, et surtout ingrate. Pour cette raison, il est vite apparu nécessaire, pour déployer des moteurs de SMT exploitables, de disposer d'une méthode d'appariement des phrases d'un corpus parallèle. C'est précisément ce qui est exposé dans Gale et Church (1993) : un algorithme d'alignement automatique des phrases dans un corpus « bilingue », c'est à dire parallèle.

### Alignement lexical

Une fois les segments-phrastiques source et cible alignés, grâce à une multitude d'heuristiques, dont : l'ordre des segments dans le corpus, leur taille, la détection d'invariants (également nommés « cognats » : noms propres, termes de forme proche, etc.), reste à parcourir

chaque paire de segments alignés, afin d'aligner les mots source avec leur(s) cible(s) potentielle(s). Afin d'induire des probabilités de transfert de la langue-source vers la langue-cible plus pertinentes, l'algorithme originel de Brown *et al*. (1988) prévoit une phase de détection de ce que les auteurs appellent « locutions fixes ». Il s'agit de segments récurrents, présentant une affinité lexicale notable, détectable automatiquement à partir de l'observation de plusieurs millions de mots du corpus de traduction. Le problème de l'appariement de ces « locutions fixes » est posé essentiellement en termes de pertinence : l'objectif est d'assurer que les segments les plus fixes de la langue-source pourront être traduits par des équivalents en langue-cible, en présupposant que les affinités lexicales auront été globalement préservées dans le corpus de traductions utilisé.

Afin de constituer le « glossaire des locutions fixes », c'est à dire la liste des paires ou triplets de mots présentant une affinité lexicale particulière, les auteurs ont une fois de plus recours à la théorie de l'information, en calculant un score d'information mutuelle entre mots d'une « locution » potentielle. Pour ce faire, tant le texte-source que le texte-cible doivent être transformés en n-grammes, c'est à dire des compositions de *n* mots contigus, avec *n* valant 1, 2, 3, ..., *i*. La transformation d'un segment traductionnel (une « phrase ») en 2-grammes est une opération élémentaire qui permet, pour chaque mot, de connaître la probabilité de transition vers le mot suivant (ou précédent). Il

devient ainsi possible d'estimer la force de l'affinité lexicale entre les deux mots d'un 2-gramme, en rapportant la fréquence de chaque mot isolé à la fréquence d'apparition de chaque mot dans la paire considérée, et donc de réaliser les propositions de King (1961). Le même principe peut s'appliquer pour des valeurs plus grandes de *n* : Brown *et al.* (1988) ont recours à des 3-grammes, notamment pour la phase finale de génération de la phrase en langue-cible. Ceci revient à estimer la probabilité d'occurrence d'une phrase entière, à partir des transitions de probabilité entre mots ($e_1$, $e_2$, … , $e_n$), c'est-à-dire très exactement ce que les tenants d'une approche mentaliste des faits linguistiques, tels que Chomsky, considèrent comme « inutile ».

### INDUIRE, PLUTÔT QUE DÉCRIRE, DES RÈGLES DE TRANSFERT

Grâce à une approche volontairement simpliste, applicable à toute paire de langue, Brown *et al.* (1988) permirent de contourner un verrou technologique majeur des systèmes de TA de l'époque : les systèmes à base de règles, bien que maîtrisables, adaptables en fonction du domaine, et déployables même en l'absence de tout corpus, sont longs à paramétrer, et surtout très complexes à faire évoluer. En effet, toute modification locale d'une règle de traitement peut avoir des effets de bord difficiles à anticiper. Or, dès la fin des années 1980, une pression accrue se fait sentir sur le monde de la traduction : la concrétisation du projet d'Union Européenne, à lui

seul, oblige les acteurs du domaine à un changement de stratégie. En effet, les moteurs de traduction étant en général développés par paires (sauf les moteurs à base d'interlangue), la simple combinatoire fait craindre une situation bloquante pour les institutions dans lesquelles des documents en grand nombre doivent désormais être traduits dans plusieurs langues. Même en ne considérant que les trois langues des membres fondateurs (allemand, français + anglais), s'il devient nécessaire de traduire la plupart des documents de l'Union Européenne (dispositions légales, statuts, débats, dispositions administratives, documents de gestion, etc.), il faut alors développer 6 modèles de traduction, puisque les règles de transfert sont ordonnées[6].

En effet, l'approche par paires de langues correspond à une situation de choix de *m* éléments parmi *n*, avec des combinaisons ordonnées. Dans notre configuration à 3 langues fondamentales correspondant aux pays membres fondateurs, la combinatoire est de 6. Le problème est bien connu : l'ajout d'un seul nouveau membre, et donc l'ajout d'une langue de traduction, fait foisonner les combinaisons. Avec 5 membres, il devient nécessaire de développer 20 modèles de traduction. En considérant une UE à 28 membres, il serait nécessaire, selon cette logique, de développer 756 modèles de traduction si chaque document devait être traduit simultanément dans toutes les langues des pays membres[7]. On le voit,

[6] En l'occurrence, français → anglais, anglais → français, français → allemand, etc.

[7] Ce calcul est réducteur : il est bien connu que les documents de l'UE ne sont pas traduits simultanément dans toutes

rien qu'avec les besoins énormes de traduction liés au fonctionnement de l'UE, sans parler d'autres institutions comme l'ONU ou le Parlement du Canada, ou encore les besoins de traduction dans les groupes privés à dimension mondiale (par exemple : Total, Thalès, etc.), le développement de moteurs de TA à base de règles, essentiellement paramétrées manuellement, aurait dû été confronté à un problème d'explosion combinatoire. Dans ce contexte, l'induction automatique de modèles de traduction, à partir de la même stratégie de base quelles que soient les paires de langues considérées, permet de répondre (en partie) au problème de la généralisation de la TA, bien qu'en pratique, on ne dispose pas de corpus parallèles dans toutes les langues concernées à ce jour.

Ainsi, Brown *et al.*, bien qu'ils ne fournissent pas dès 1988 de résultats de traduction comparables à ceux de Systran, montrent comment des règles de transfert français → anglais peuvent être acquises à partir d'un corpus parallèle (*i.e.* Hansard), en sélectionnant les candidats sur la base du score maximal d'information mutuelle (*Cf.* fig. 1).



---

les langues des pays membres.

| | |
|---|---|
| eau | *water* |
| banque | *bank* |
| votre | *your* |
| enfants | *children* |
| trop | *too* |
| aujourd'hui | *today* |
| seulement | *only* |
| peut | *cannot* |
| ceintures | *seat* |
| ceintures | *belts* |
| bravo | ! |

Fig. 1. Candidats-traduction induits de façon automatique, tirés de Brown *et al.* (1988)

Comme on peut le voir, bien que le résultat soit perfectible, les rapprochements proposés sont tous cohérents, y compris les erreurs manifestes (ceinture / *seat*), dues soit à la nature des unités rapprochées (un mot simple en français / un mot composé en anglais), soit à la nature du corpus (bravo / !).

## LA PBMT, UNE EXTENSION DE LA SMT

Nous l'avons vu plus haut : dès la fin des années 1980, la demande accrue en traduction (humaine comme automatique), associée à

l’accumulation d’exemples de traduction liée à la vie d’institutions politiques multilingues (ONU, UE, etc.) rendent possible, voire nécessaire, un changement de paradigme, des systèmes à base de règles, guidés par les connaissances, vers des systèmes paramétrés à partir d’exemples concrets de traductions. Toutefois, comme le rappelle Wilks (2008), les performances des premiers moteurs de SMT étaient loin d’égaler celles de Systran, pour la paire français → anglais, tout du moins : au mieux les nouveaux systèmes réussissaient-ils à traduire convenablement environ 50 % des phrases qui leur étaient soumises. À partir des premiers succès enregistrés par les moteurs de SMT, de nouvelles stratégies furent explorées, afin d’améliorer notamment le rendu en langue-cible, par exemple en repérant de façon plus fine et plus systématique les séquences récurrentes : unités polylexicales, langage formulaïque, unités phraséologiques, termes techniques composés, etc.

C’est dans ce contexte qu’émergent les moteurs de traduction à base de séquences, ou *Phrase-Based Machine Translation*. Ces systèmes constituent un raffinement des moteurs de SMT : ces nouveaux systèmes visent explicitement à identifier des correspondances entre séquences de mots, et non mots isolés, entre la source et la cible. En ce sens, les moteurs PBMT cherchent à introduire davantage d’informations contextuelles dans le processus de transfert de la source vers la cible.

## La prise en charge des associations lexicales

Koehn, Och *et al.* (2003) présentent l'un des premiers moteurs de PBMT, pensé comme une extension de la SMT. Les auteurs soulignent le fait que l'exploitation des *phrases*, ou séquences-pivots, était déjà présente dans des réalisations antérieures : Och, Tillmann *et al.* (1999) et Marcu et Wong (2002), du côté des systèmes SMT, ainsi que Yamada et Knight (2001) du côté des stratégies alliant approche statistique et représentations syntaxiques formalisées, pour la traduction de langues typologiquement éloignées (par exemple : anglais/japonais). Koehn, Och et al. (2003) insistent sur le fait que de bonnes performances en traduction peuvent être atteintes en ayant recours à des principes simples, voire que les approches plus sophistiquées (analyse syntaxique) dégradent les performances. On le voit, l'objectif des auteurs est avant tout d'améliorer l'efficacité d'un système de type SMT, en faisant le pari d'une approche linguistiquement pauvre en connaissances : *small phrases of up to three words are sufficient for obtaining high levels of accuracy* (« des séquences de petite taille, jusqu'à trois mots, sont suffisantes pour obtenir des scores de précision élevés »). Dans cette approche, les phases d'alignement de phrases, puis de mots, déterminent largement les résultats finaux. Toutefois, les paramètres de chaque modèle de traduction doivent être définis de façon empirique, sans principe théorique général. L'objectif de performance prime donc sur la compréhension du système linguistique, ou toute

autre hypothèse traductologique. L'approche repose essentiellement sur des heuristiques, dont la plus simple (extraction de séquences par exploration des alignements lexicaux) se révèle également la plus efficace.

Le développement des moteurs de PBMT correspond à l'intuition que de nombreuses séquences de mots sont répétées quasiment à l'identique, tant en langue-source qu'en langue-cible : formules toutes faites, *phrasal verbs* de l'anglais, locutions telles que *right on top of* (« juste au-dessus de »), expressions idiomatiques telles que *to bring the house down* (« casser la baraque »), voire proverbes ou plus largement collocations, *lexical bundles* (« paquets lexicaux ») de Biber *et al.* (2004) et autres unités polylexicales (noms propres, dates, termes complexes, etc.).

## De l'alignement de mots à l'alignement de séquences

Les systèmes PBMT reprennent l'architecture générale des moteurs de SMT. D'après Koehn, Och *et al.* (2003), à partir du modèle du canal bruité, l'application du théorème de Bayes permet de concevoir la traduction comme suit : étant donnée une phrase en langue-source, la traduction d'une phrase vers la langue-cible (en l'occurrence l'anglais) est donnée par la sélection de la traduction présentant la probabilité la plus élevée (parmi celles observées). En formalisant le problème de cette façon, il suffit de disposer d'un modèle de la langue voulue (ici *e* pour l'anglais) et d'un modèle de traduction *p(f|e)* (probabilité

d'avoir *f* sachant *e*). En considérant des séquences-pivots de longueur fixe (des n-grammes, avec *n* valant de 1 à 6 en général), et non plus des phrases complètes, et de procédant à une phase de réordonnancement des mots, on obtient une traduction automatique qui combine les avantages de la SMT classique avec (une partie de) ceux des approches reposant sur des séquences linguistiquement bien formées (ex. : traduction par transfert syntaxique).

L'élément-clé dans un système de PBMT est l'algorithme d'alignement lexical : l'approche préconisée par Koehn, Och *et al.* (2003) enregistre les meilleures performances pour une extraction de séquences-pivot à partir des tables d'alignement lexical produites par la plate-forme Giza++, détaillée dans Koehn, Hoang *et al.* (2007)[8]. Au cours de la phase d'alignement, les séquences contenant des mots, eux-mêmes réputés alignés, sont à leur tour alignées. Les probabilités de traduction d'une séquence-source vers une séquence-cible guident la constitution d'une table des séquences-pivots, qui s'ajoute à la table des alignements lexicaux.

L'un des résultats les plus marquants rapportés par Koehn, Och *et al.* (2003) est que l'approche la plus simple, qui repose directement sur les alignements lexicaux disponibles dans tous les moteurs de SMT, est celle qui donne les meilleurs résultats. Par ailleurs, contrairement à ce qu'un linguiste pourrait espérer, laisser le système aligner les séquences les plus longues

[8] Disponible sur http://www.statmt.org/moses/giza/GIZA++.html

n’apporte qu’un gain modique : le meilleur compromis semble être de limiter la taille des séquences à des 3-grammes, d’après Koen *et al*. Les performances continuent de s’améliorer jusqu’au niveau des 7-grammes, toutefois l’investissement en temps de traitement et en consommation de mémoire semble peu rentable, par rapport aux performances déjà quasiment maximales obtenues avec des 3-grammes. Dans l’évaluation menée par les auteurs, les pires résultats sont enregistrés par les approches contraignant la bonne formation syntaxique des séquences extraites, dans une optique de traduction par transfert syntaxique, et non simplement lexical.

## DISCUSSION DE QUELQUES EXEMPLES

Après avoir présenté les grandes lignes de l’approche PBMT, nous nous arrêtons ici sur des exemples que nous jugeons illustratifs du fonctionnement de ces systèmes, tant dans leurs atouts que dans leurs lacunes.

### SÉQUENCES INDUITES DES CORPUS EUROPARL

Les exemples discutés ici sont identifiés de façon automatique par un moteur de PBMT, en l’occurrence la plate-forme MOSES[9], présentée

[9] Disponible au téléchargement sur http://www.statmt.org/moses

dans Koehn, Hoang *et al.* (2007). Les données utilisées ici sont les tables de séquences (*phrase tables*) induites à partir des corpus parallèles Europarl[10].

Comme nous l'avons vu plus haut, l'une des étapes primordiales pour un moteur de SMT consiste en l'identification des potentiels alignements de mots, pour chaque segment phrastique aligné automatiquement entre la source et la cible. Dans le cas de la plate-forme Moses, lors du prétraitement des corpus parallèles (en l'occurrence les corpus Europarl anglais/français, ici), l'étape de constitution des tables de mots fournit des listes telles que celle présentée dans la fig. 2[11], pour des traductions anglais → français.



[10] Fichier compressé phrase-table.1.gz (date de création 2017-07-26 05:37, 2,7 Go) fourni par les développeurs de la plate-forme Moses (version 4.0).

[11] Le tableau reprend les données du fichier lex.1.e2f, disponible à la même adresse que les autres fichiers mentionnés précédemment.

| Mot | Candidat-traduction | Probabilité |
|---|---|---|
| *derogatory* | dénigrants | 0.7500000 |
| - | péjoratives | 0.3333333 |
| - | désobligeants | 0.3333333 |
| - | mériterions | 0.2500000 |
| - | utilisais | 0.2000000 |
| - | désobligeante | 0.2000000 |
| - | désobligeant | 0.1428571 |
| - | désobligeantes | 0.1333333 |
| - | dénigrant | 0.1000000 |
| - | dénigre | 0.0666667 |
| - | péjorative | 0.0588235 |
| - | dérogatoire | 0.0564516 |
| - | péjoratif | 0.0526316 |
| - | méprisante | 0.0400000 |
| - | dégradantes | 0.0277778 |
| - | proférer | 0.0277777 |

| | | |
|---|---|---|
| | | 8 |

Fig. 2 . Quelques candidats-traduction pour *derogatory*

Nous voyons ici un extrait de la liste des candidats traduction pour *derogatory*, triés selon leur probabilité calculée en corpus. Le traitement automatique montre bien la polysémie du terme anglais, pouvant se rapporter aussi bien à des propos jugés insultants que, dans le domaine juridique, à une clause dérogatoire. Des scories, causées par des rapprochements indus entre le terme-source (un adjectif) et les candidats, sont toutefois visibles. Notamment, le système rapproche *derogatory* de mots grammaticaux[12] (préposition, déterminants, pronoms), ainsi que de catégories qui, bien que cooccurrentes du terme-cible, ne seraient pas envisagées par un traducteur humain. C'est le cas notamment de « mériterions », « utilisais », et « proférer ». Par la suite, la procédure d'identification de séquences cherche, pour chaque alignement lexical dont le score est suffisamment élevé, les cooccurrents les plus longs et/ou ceux présentant un score supérieur à un seuil donné. Cette procédure, à partir des candidats-traduction identifiés pour *derogatory*, produit des appariements tels que présentés dans la fig. 3.

[12] Non présentés dans la fig. 2 pour des raisons de place.

| **Séquence-source** | **Candidat-traduction** |
|---|---|
| *a derogatory Russian term* | une insulte russe |
| *denoting that* | qui fait de |
| *derogatory views* | des avis désobligeants |
| *extremely derogatory* | extrêmement désobligeants |
| *and insulting* | et insultants |
| *B and C penalise* | B et C criminalisent |
| *derogatory remarks* | les remarques péjoratives |
| *a derogatory regime* | un régime dérogatoire |
| *and the derogatory nature and* | ainsi que le caractère dérogatoire et |
| *any derogatory way* | importe quelle façon désobligeante |
| *such derogatory terms .* | proférer . |

Fig. 3. Alignement de séquences construites sur *derogatory*

Comme on peut le voir dans la fig. 3, la traduction par séquences-pivots permet d'exploiter les régularités de cooccurrences entre unités lexicales. Ainsi, non seulement le système distingue, de façon purement mécanique, les différents emplois de *derogatory*, mais il distingue également les différentes acceptions d'autres mots polysémiques, qui cooccurrent avec le mot recherché. C'est le cas de *regime*, qui, comme en français, peut désigner aussi bien un système politique qu'une disposition légale, entre autres.

Bien que, là encore, des scories soient présentes, le mécanisme d'appariement de séquences propose des rapprochements qui, pour un traducteur, sont loin d'être inintéressants. Bien sûr, la procédure étant, de l'aveu même de ses concepteurs, non linguistique, les séquences ainsi extraites ne présentent aucune unité, notamment en termes de construction syntaxique. De ce fait, de la même façon que des rapprochements indus étaient proposés au niveau lexical (*Cf.* fig. 3), des séquences telles que « proférer . » sont rapprochées de *such derogatory terms .* alors qu'en langue-source, la séquence constitue un syntagme nominal, et qu'en langue-cible, elle s'articule autour d'un verbe (cooccurrent habituel de « insultes », toutefois).

Nous n'examinerons pas en détail les occurrences particulières qui occasionnent des rapprochements indus. Nous nous bornerons à constater que la plupart des rapprochements proposés paraissent pertinents, bien que les tables de séquences ainsi constituées soient redondantes et, dans le même temps, lacunaires : de nombreuses variantes du même patron fondamental de construction sont considérées, alors que, par ailleurs, des séquences telles que « propos désobligeants », voire « injure » sont tout simplement absents[13]. Signalons ici une autre propriété fondamentale des moteurs de traduction PBMT : leur capacité à détecter des motifs récurrents, qui, en fonction des textes, peuvent

[13] La série des noms se rapportant à des verbes de parole est cependant bien représentée : *termes*, *avis*, *remarque*, *expression* et *commentaires* sont ainsi présents.

représenter une part non négligeable du texte à traduire. C'est le cas, entre autres, d'expressions de quantité telles que *$ 1 000 for*, *$ 1 000* ou *$ 1 billion in the*, automatiquement alignées avec leur correspondance en langue-cible, respectivement : « 1000 dollars pour », « 1000 dollars » et « milliard de dollars ». Bien qu'on puisse se réjouir que l'outil automatique soit à même de repérer des régularités de construction[14], il n'en reste pas moins que la procédure produit toujours des séquences qui peuvent s'avérer problématiques, telles que par exemple : *$1 billion*/« élève à environ 1 milliard de dollars[15] ».

## De vraies unités phraséologiques parmi de simples « unités-pivots »

Les concepteurs des moteurs de traduction PBMT revendiquent le fait que les *phrases* extraites des corpus n'ont aucun statut linguistique. Ces séquences ne seraient ainsi que de simples pivots améliorés, utiles pour le transfert de chaînes de caractères de la langue-source vers la langue-cible. Après des décennies d'une suprématie incontestée de moteurs de TA à base de règles, cette prise de position peut sembler quelque peu surprenante. L'une des raisons derrière cette stratégie volontairement pauvre en connaissances linguistiques tient au peu de précision des analyses syntaxiques automatiques disponibles : bien que les analyseurs

[14] Ce qui laisse espérer au traducteur humain de déléguer de façon efficace une partie non négligeable du travail de traduction.

[15] Toutefois, si ces séquences ont été repérées par l'algorithme, c'est bien parce que l'une est souvent proposée en traduction pour l'autre.

syntaxiques aient, eux aussi, opéré un changement de paradigme des systèmes guidés par les connaissances vers des systèmes guidés par les données (des corpus de type Treebank), il faut bien admettre que même les analyseurs les plus doués peinent, encore aujourd'hui, à distinguer un simple complément d'objet indirect d'un ajout prépositionnel. Par ailleurs, le développement de moteurs de traduction guidés par les données obéit justement à une volonté de s'abstraire des langues particulières, et donc des choix d'analyse propres aux paires de langues considérées. Dans ces systèmes, l'accent est donc mis sur l'optimisation des modèles statistiques induits à partir de corpus : modèles de langue (probabilités de transitions entre mots), modèles de la distorsion du message, modèles du réagencement des traductions en langue-cible. Toutefois, à partir du moment où ces systèmes brassent des millions d'exemples de traductions, on peut supposer que, parmi la multitude de simples séquences-pivot se trouvent au moins quelques unités phraséologiques, ou à tout le moins, des unités polylexicales, notoirement difficiles à traduire. Nous examinons ci-dessous deux familles de structures : la première (« être/se mettre au diapason ») peut être caractérisée comme une expression idiomatique. La seconde (*in view of*) relève plutôt du *lexical bundle* de Biber : une séquence récurrente de mots, propres à un domaine d'activité ou à un registre de langue.

### « ETRE/SE METTRE AU DIAPASON »

Le thème des prises de décision collectives, de l'harmonisation nécessaire des points de vue (par exemple : les attentes de la population et les possibilités de l'appareil politique) est abondamment représenté dans le corpus Europarl. Nous nous penchons ici sur les structures anglaises rendues en français par l'expression idiomatique « être/se mettre au diapason ». En effet, il n'existe pas d'équivalent direct de l'expression française, ces formes sont donc nécessairement le résultat de l'interprétation de situations impliquant un processus tel que :

- l'harmonisation des idées, du rythme ou de la direction de travail ;
- l'adaptation d'un individu à un contexte changeant ;
- par extension, l'adaptation d'un collectif (institution, entreprise) à un contexte changeant.

Comme on peut le voir sur la fig. 4, dans la plupart des cas, l'expression idiomatique française rend une tournure elle-même idiomatique en anglais. Toutefois, il semble que l'expression française soit également fréquemment observée dans des contextes non idiomatiques en langue-source, comme par exemple : *adjust to its surroundings* (« s'adapter à son environnement ») ou *match its financial status* (« correspond(re) à son statut financier »). L'expression idiomatique française est par ailleurs exploitée pour rendre des tournures idiomatiques très différentes, telles que *X is going Y* (« X se lance dans (le domaine)

Y ») et *to put one's money where one's mouth* is (« joindre l'acte à la parole »).

| Séquence-source | Candidat-traduction |
|---|---|
| *Commission are pulling together* | Commission sont au diapason |
| *European institutions fall into step and* | institutions européennes mettent au diapason et fassent |
| *Member States singing from the same song* | au diapason des États membres |
| *aligning itself with the mainstream of events* | se mettre au diapason des événements |
| *are not keeping in step with the* | ne sont pas au diapason avec les |
| *be in line with this .* | soyons au diapason de cela . |
| *keeping in step with the wishes* | au diapason avec les souhaits |
| *match its financial status as* | au diapason de son statut financier de |
| *to keep up with the trend .* | aussi se mettre au diapason . |
| *make sure they stay abreast of new* | se mettre au diapason des |
| *tune with the natural environment ,* | au diapason du milieu naturel , de |

Fig. 4. Quelques traductions automatiques d'une expression idiomatique française

Cette mise en correspondance automatique d'un ensemble de constructions très diverses en

langue-source (anglais), avec une expression idiomatique en langue-cible (français) qui joue sur la métaphore vive de l'orchestre qui s'accorde au même référentiel, révèle des choix de traduction qui, en l'absence de métadonnées complètes, ne peuvent être qu'imparfaitement reconstruits[16]. L'expression « être/se mettre au diapason » semble jouer le rôle d'expression par défaut, voire de cliché. En effet, d'autres traductions étaient possibles pour les différentes constructions anglaises concernées. Il est toutefois difficile de déterminer si ce biais est le résultat de choix stylistiques individuels et inconscients de la part des traducteurs humains, ou de conventions stylistiques explicites (charte éditoriale) au sein de l'institution.

### IN VIEW OF

Examinons à présent le segment *in view of*. Cette séquence assez banale[17], qui marque cependant un passage argumentatif dans un document officiel, pourrait être transposée directement en français par « au vu de ». Toutefois, l'examen des tables de correspondances produites par Moses révèle, là encore, des choix traductionnels particuliers. En effet, loin d'être simplement transposé en français, le paquet

[16] Les documents composant le corpus Europarl ne sont pas tous complets, en termes de métadonnées : il est parfois difficile de déterminer dans quelle langue s'exprimait l'orateur, et si un document donné est à l'origine en anglais ou non. En tout état de cause, ces métadonnées sont perdues lors de la phase d'identification des séquences-pivots.

[17] Et fréquente : plus de 9 600 occurrences dans la seule table discutée ici.

lexical *in view of* semble devoir être constamment explicité par les traducteurs. Les choix de traductions peuvent être regroupés comme suit :

- traductions quasi-littérales (dimension visuelle) : « au vu de », « en vue de », « vu », « en regard de », « au regard de », « à la lumière de » ;
- traductions adaptées (dimension visuelle absente) : « afin de », « à la suite de », « après », « car », « compte tenu », « concernant », « dans la perspective de », « dans le cadre de », « dans l'optique », « de par », « devant », « dû », « en considérant », « en prévision de », « en raison de », « en tenant compte de », « étant donné », « eu égard à », « face à », « par », « parce que », « par rapport à », « partant », « pour », « prenant en considération », « sachant », « suite à », « visant à » ;
- explicitation de sens : « en entendant cela », « devant », « compte tenu/en tenant compte de », « étant donné ».

L'examen des traductions différentes pour une même expression, qui aurait très bien pu être transposée directement en français, révèle différentes dimensions sémantiques. La première dimension qui s'impose est celle du lien causal, vu soit comme la cause déclenchante, soit comme une finalité ou un but à atteindre. La cause déclenchante semble constituer un élément d'argumentation pour expliquer, *post hoc*, une décision ou une situation. Enfin, un dernier cas de lien causal se révèle : celui d'un lien ténu, dans

lequel aucune orientation causale ne semble réellement dominante. En considérant le lien causal, les choix de traduction français se regroupent comme suit :

- finalité/but : « afin de », « car », « dans l'optique », « dans la perspective de », « devant », « en vue de », « face à », « pour », « par rapport à », « visant à » ;
- explication *post-hoc* : « à la lumière de », « à la suite de », « après », « au regard de », « au vu de », « compte tenu », « de par », « dû », « en considérant », « en raison de », « en regard de », « en tenant compte de », « étant donné », « eu égard à », « face à », « par », « parce que », « partant », « prenant en considération », « sachant », « suite à », « vu » ;
- lien de causalité ténu : « concernant », « dans le cadre de », « par rapport à ».

On le voit, la séquence *in view of* donne lieu en français à de nombreuses adaptations, selon la dimension sémantico-discursive favorisée par le traducteur. Signalons un absent, toutefois : « en préparation de » était attendu et ne semble pas attesté dans la table de séquences induites du corpus examinée ici.

## SYNTHÈSE ET PERSPECTIVES

## LA NMT A-T-ELLE TUÉ LA PBMT ?

Les premiers moteurs de PBMT remontent, d'après Koehn, Och *et al.* (2003) à la fin des années 1990. Toutefois, il faut attendre 2003, pour que la stratégie de traduction par séquences-pivots soit largement diffusée, grâce à la disponibilité de la plate-forme Moses. Autant dire une éternité, du point de vue de l'évolution technologique en intelligence artificielle. Or, les géants du web, et en particulier les GAFAM (Google, Amazon, Facebook, Apple et Microsoft) ont investi le terrain de la TA, et ce depuis plusieurs décennies dans le cas de Microsoft[18]. Au cours des cinq dernières années, Google, talonné par Facebook, ainsi que de nouveaux arrivants tels que DeepL, ont réalisé des percées technologiques notables, en adoptant des plateformes d'apprentissage automatique reposant sur des architectures de plus en plus complexes de réseaux de neurones artificiels. Les acteurs historiques du domaine, tels que Systran, semblent avoir définitivement perdu leur position hégémonique. Par ailleurs, tant Google que Facebook proposent en accès libre leur plateforme d'apprentissage machine, ainsi que des modèles de traduction induits à partir de corpus parallèles de type Europarl. L'objectif est clair : convaincre les acteurs du marché que la NMT est la technologie d'avenir en TA, tout en s'assurant de la domination technologique par l'adhésion des développeurs à leurs produits, respectivement

[18] Il manque ici les équipes d'IBM, sans lesquelles la SMT n'aurait jamais été remise au goût du jour.

TensorFlow, pour Google, et Unsupervised MT pour Facebook. Les deux géants du web visent à brève échéance le contournement de l'un des goulots d'étranglement les plus importants : les équipes de Google ont récemment annoncé la possibilité de réaliser des traductions par associativité, notamment dans le cas des langues sous-dotées. Quant à Facebook AI Research, ses équipes décrivent dans Lample *et al.* (2018) des modèles hybrides, alliant les avantages de la PBMT (précision, respect des contraintes structurelles, volumes de corpus de paramétrage limités) à ceux de la NMT (capacité de généralisation, voire traduction par représentations « profondes »), non supervisée de surcroît (possibilité de paramétrer un moteur de TA sur un corpus monolingue).

Malgré des avancées marquantes, la traduction neuronale semble donc loin d'avoir tué la PBMT, bien au contraire. Comme le prédisait Wilks (2008), le temps semble venu pour les approches hybrides, cumulant les avantages de chaque paradigme, y compris, pourquoi pas, les approches par règles lorsqu'elles apportent un gain mesurable.

## DIMENSION HEURISTIQUE DES PHRASES

À ce stade, il nous paraît nécessaire de revenir sur le statut des séquences-pivots que sont les *phrases* de la PBMT. Le recours à ces séquences a, dès le départ, été conçu en-dehors, voire en opposition, à toute approche linguistiquement fondée. Du point de vue du linguiste, les succès

enregistrés ces dernières années par les approches probabilistes en TAL, au-delà de la TA (example : reconnaissance vocale), soulignent à nouveau la nécessité de dépasser la dichotomie Compétence/Performance (ou Langue/Parole), tant dans les domaines appliqués que dans les domaines plus théoriques. En effet, ce qu'interrogent les succès de la SMT, et en particulier de la PBMT, c'est justement le statut de la structure syntaxique pour la traduction, mais également au-delà. D'après Wilks (*ibid.*), le tournant empirique de la TA peut être vu comme un changement de paradigme aussi important que l'abandon du mimétisme biologique dans la conception de machines volantes, en aéronautique. Toutefois, la question reste posée de la portée théorique, voire épistémologique, des succès remportés par les approches empiriques en TA.

## L'IVROGNE ET LE RÉVERBÈRE

Dans Habert (2005), l'auteur convoque la figure de l'ivrogne, cherchant ses clés la nuit, sous un réverbère, non parce que c'est là qu'il les a perdues, mais parce qu'au moins, à cet endroit, on y voit clair. Pour Habert, les linguistes disposent désormais d'outils informatiques génériques (par exemple : bases de données) et d'instruments dédiés (par exemple : étiqueteurs en parties du discours, programmes d'alignement des segments phrastiques d'un corpus parallèle) qui leur permettent de modéliser de façon exhaustive et vérifiable des faits linguistiques, et donc de les

traiter de façon automatique. Ces outils et instruments, à partir du moment où ils sont constitués comme dispositifs, permettent également de tester des hypothèses. Malgré tout, pour Habert, les linguistes outillés sont souvent réduits à n'observer qu'une partie infime du territoire : celle qui est accessible aux outils et instruments, non parce qu'elle recèle des faits pertinents, mais bien parce qu'elle est « éclairée ».

Nous avons montré plus haut comment des séquences extraites de façon aveugle, par simple exploration d'alignements lexicaux, pouvaient mettre en lumière des dimensions sémantiques, voire discursives, sous-tendant les choix de traduction. Toutefois, nous l'avons également montré, bien que ces séquences-pivots présentent un intérêt manifeste pour le traducteur, notamment dans le cadre d'une chaîne éditoriale intégrant des aides à la traduction (mémoire de traduction), il n'en reste pas moins que les séquences extraites par un moteur de PBMT sont conformes au programme annoncé par ses concepteurs : elles ne revendiquent aucun statut linguistique. Elles sont donc redondantes, et malgré tout lacunaires, comme nous avons pu le montrer à partir de quelques exemples. Afin de conférer à ces séquences-pivots un statut proprement linguistique, il serait nécessaire, par exemple, de leur appliquer les algorithmes d'inférence de segment traductionnels des moteurs de traduction par analogie. Il serait alors possible de constituer un relevé exhaustif de motifs plus longs, voire discontinus, et de pallier

en partie la cécité linguistique de l'approche PBMT telle qu'elle est implémentée à l'heure actuelle.

Toutes les séquences-pivots extraites d'un corpus par un moteur PBMT ne révèlent pas, loin s'en faut, des faits linguistiques ou traductologiques pertinents. Toutefois, nous avons essayé de montrer comment ces tables de séquences pouvaient constituer l'équivalent d'un filon potentiellement riche, au sein de la mine à ciel ouvert que sont les données massives d'un corpus parallèle, pour un linguiste traquant les unités polylexicales et les unités phraséologiques. Ces tables peuvent ainsi être vues comme des super-concordanciers, ciblant explicitement les segments récurrents des langues en présence, pour en fournir une liste exhaustive. Ainsi, même si les moteurs de PBMT semblent pour l'heure techniquement dépassés par la NMT, ils n'en constituent pas moins un outil (au sens de Habert) irremplaçable pour l'étude des unités polylexicales et de la phraséologie en corpus. Nous espérons avoir également montré que cet outil pouvait être amené à devenir un véritable instrument d'exploration des choix de traduction, voire un dispositif expérimental pour la traductologie (identification d'unités phraséologiques), à condition d'y adjoindre un mécanisme de généralisation des séquences-pivots. Ce faisant, nous espérons avoir évité le double écueil de l'ivrogne au réverbère : ne chercher que là où on y voit clair, mais également s'aider du réverbère plus parce qu'il soutient que parce qu'il éclaire.

Antonio Balvet
Université de Lille – SHS, UMR 8163 STL
France

RÉFÉRENCES BIBLIOGRAPHIQUES

BIBER, Douglas, CONRAD, Susan, CORTES, Viviana, 2004, « If you look at... : Lexical bundles in university teaching and textbooks », *Applied linguistics,* vol. 25, n° 3, p. 371-405.

BROWN, Peter, COCKE, John, DELLA PIETRA, Stephen, DELLA PIETRA, Vincent, JELINEK, Frederick, MERCER, Robert, ROOSSIN, Paul, 1988, « A statistical approach to language translation », in *Proceedings of the 12th conference on Computational linguistics*, vol. 1, Association for Computational Linguistics, p. 71-76.

CHOMSKY, Noam, *Syntactic structure*, Mouton, 1975.

CHOMSKY, Noam, « Quine's empirical assumptions », *Words and Objections: Essays on the Work of W.V. Quine*, (dir.) Davidson, Donald, Hintikka, Jaakko, D. Reidel, Dordrecht, 1969, p. 53-68.

DE LAPLACE, Simon P., *Essai philosophique sur les probabilités*, 1840, Gauthier-Villars.

GALE, William A., CHURCH, Kenneth W., 1993, « A program for aligning sentences in bilingual corpora », *Computational linguistics*, vol. 19, n° 1, p. 75-102.

GOODFELLOW, Ian, BENGIO, Yoshua, COURVILLE, Aaron, 2016, *Deep learning*, t. 1MIT press Cambridge, 2016, t. 1.

HABERT, Benoît, « Portrait de linguiste(s) à l'instrument », *Revue Texto,* vol. 10, n° 4, 2005.

JOHNSON, Melvin, SCHUSTER, Mike , LE, Quoc V, KRIKUN, Maxim, WU, Yonghui, CHEN, Zhifeng, THORAT, Nikhil, VIÉGAS, Fernanda, WATTENBERG, Martin, CORRADO, Greg *et al.,* « Google's multilingual neural machine translation system: enabling zero-shot translation », 2016, arXiv preprint, arXiv :1611.04558.

KING, Gilbert W., « Stochastic methods of mechanical translation », in *Readings in machine translation*, MIT Press Cambridge, 1961, p. 45-51.

KOEHN, Philipp, HOANG, Hieu, BIRCH, Alexandra, CALLISON-BURCH, Chris, FEDERICO, Marcello, BERTOLDI, Nicola, COWAN, Brooke, SHEN, Wade, MORAN, Christine, ZENS, Richard *et al.,* « Moses: Open source toolkit for statistical machine translation », *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, *Association for Computational Linguistics*, 2007, p. 177-180.

KOEHN, Philipp, OCH, Franz Josef, MARCU, Daniel, « Statistical phrase-based translation », *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language*

*Technology*, vol. 1, *Association for Computational Linguistics*, 2003, p. 48-54.

LAMPLE, Guillaume, OTT, Myle, CONNEAU, Alexis, DENOYER, Ludovic, RANZATO, Marc'Aurelio, « Phrase-Based & Neural Unsupervised Machine Translation », 2018, arXiv preprint arXiv :1804.07755.

MARCU, Daniel, WONG, William, « A phrase-based, joint probability model for statistical machine translation », *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, vol. 10, *Association for Computational Linguistics*, 2002, p. 133-139.

NAGAO, Makoto, « A framework of a mechanical translation between Japanese and English by analogy principle », *Artificial and human intelligence*, 1984, p. 351-354.

OCH, Franz Josef & NEY, Hermann, « A Systematic Comparison of Various Statistical Alignment Models », *Computational Linguistics*, vol. 29, n° 1, 2003, p. 19-51.

OCH, Franz Josef, TILLMANN, Christoph, NEY, Hermann, « Improved alignment models for statistical machine translation », *1999 Joint SIG-DAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora,* 1999.

SHANNON, Claude E., « A mathematical theory of communication », *Bell Systems Technology Journal,* vol. 27, 1948, p. 623-656.

WILKS, Yorick, *Machine translation: its scope and limits*, Springer Science & Business Media, 2008.

YAMADA, Kenji, KNIGHT, Kevin, « A syntax-based statistical translation model », *Proceedings of the 39th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics*, 2001, p. 523-530.

WEAVER, Warren, « The mathematics of communication », *Scientific American,* vol. 181, 1949, p. 11-15.