Traduire sans comprendre ? La place de la sémantique en traduction automatique

Thierry Poibeau

Traduire sans comprendre ?
La place de la sémantique en traduction automatique

Thierry Poibeau

Laboratoire LATTICE
CNRS & Ecole normale supérieure & Université Sorbonne Nouvelle
1, rue Maurice Arnoux
92120 Montrouge

Résumé

Une bonne traduction implique de comprendre le texte à traduire pour le transposer aussi finement que possible dans une langue cible. La traduction automatique a longtemps reposé sur ce constat et c'est essentiellement par une absence ou du moins une déficience au niveau de l'analyse sémantique que l'on a longtemps expliqué l'échec de la première vague de systèmes de traduction automatique (1950-1965). Les recherches se sont alors orientées vers la « compréhension automatique » visant justement à fournir une représentation formelle du contenu d'un texte. A rebours de cette conception quasi unanimement acceptée à la fin des années 1980, on a vu se généraliser depuis les années 1990 des systèmes reposant sur une approche statistique et sans représentation explicite de la sémantique des textes. On s'interrogera sur cet état de choses, sur la raison du succès des systèmes actuels et sur leurs limites éventuelles. En particulier, on examinera le rapport entre le type de traitement utilisé par les systèmes actuels et la sémantique : en partant de données brutes, attestées, en grande largeur, les systèmes n'intègrent-ils pas finalement, et contrairement à ce que pourrait laisser penser un examen rapide de ces systèmes, des connaissances importantes de nature sémantique (comme par exemple un découpage implicite suivant les différents sens du mot reposant sur une analyse distributionnelle ou une modélisation indirecte du degré de figement des expressions complexes) ?

Mots clés

Traduction automatique ; sémantique ; évaluation

1 Introduction

On sait que la traduction est un des plus vieux rêves de l'Intelligence Artificielle. Ce n'est pas un hasard : par delà les intérêts commerciaux actuels, les enjeux sont fondamentaux pour comprendre les processus cognitifs impliqués dans la communication, la compréhension, et plus généralement la faculté de langage dans toute sa complexité. C'est pourquoi dès les débuts de l'informatique, avant même la naissance des ordinateurs, des chercheurs de différents horizons se sont intéressés à ce problème.

Pour ce qui concerne la période récente, on peut schématiquement distinguer deux types de systèmes. D'une part des systèmes fondés sur des dictionnaires et des

règles d'analyse puis de transfert d'une langue à l'autre. D'autre part, des systèmes statistiques s'appuyant sur des corpus bilingues alignés, n'utilisant pas de ressources spécifiques aux langues visées. Cette description est évidemment sommaire : la plupart des systèmes sont aujourd'hui hybrides, dans la mesure où ils mélangent approche statistique et connaissances prédéfinies sur la langue. Il n'empêche, la distinction entre deux types de systèmes très différents garde en partie sa légitimité et permet d'aborder la question de la sémantique dans le domaine de la traduction automatique.

La traduction automatique statistique est vue, à juste titre, comme un domaine complexe et très technique, qui se fonde essentiellement sur le traitement de grandes masses de données et laisse à l'arrière-plan l'analyse linguistique traditionnelle. De fait, les linguistes s'y sont peu intéressés, d'autant qu'un examen rapide et superficiel suffit à montrer que ces systèmes sont encore loin d'être parfaits et qu'il est facile de les mettre en défaut. On ne peut cependant nier que, depuis une décennie ou deux, des progrès évidents ont été enregistrés et ce particulièrement au niveau de la recherche des équivalents traductionnels (c'est-à-dire, pour simplifier, au niveau de l'analyse lexicale). C'est sur ce point que nous souhaitons nous interroger dans ce qui suit : des traitements opérant directement sur de grands corpus bruts, sans connaissance *a priori* et sans ressources extérieures, peuvent-ils être efficaces ? Si oui, pourquoi ? Quelle est la nature des informations ainsi modélisées ? Peut-on les qualifier de « sémantique » ?

L'article se compose de deux parties principales. La première donne un aperçu rapide de l'histoire du domaine et permet ainsi d'examiner la place de la sémantique dans le processus. La seconde partie examine plus précisément la nature des informations modélisées par les systèmes actuels à base statistique, permettant d'une certaine façon de voir les bienfaits d'une analyse purement distributionnelle mais aussi les problèmes de l'approche quand on s'éloigne de l'anglais. L'anglais a en effet un double avantage : c'est la langue la plus répandue sur Internet et c'est surtout une langue très pauvre sur le plan morphologique, un élément crucial pour les approches aujourd'hui en vogue. On conclut enfin en rappelant que d'autres défis, au-delà du niveau lexical, sont encore loin d'être résolus malgré leur caractère fondamental pour la traduction automatique.

2 Deux approches opposées en traduction automatique

Dans cette section, nous présentons les deux approches principales en matière de traduction automatique, et la place de la sémantique dans chacune de ces approches. Comme on l'a déjà dit, la dichotomie opérée entre systèmes à base linguistique et systèmes à base statistique est un peu artificielle (les systèmes sont aujourd'hui pour la plupart « hybrides ») mais permet, nous semble-t-il, de bien cerner la problématique.

2.1 Les systèmes à base de règles

L'histoire des débuts de la traduction automatique est bien connue, notamment grâce aux travaux de John Hutchins qui a rédigé des livres et des notices sur la plupart des acteurs de ce domaine (Hutchins, 1986, 1992 ; voir aussi Nirenburg *et al.*, 2003).

Les débuts de la traduction automatique, juste après la seconde guerre mondiale, sont marqués par deux approches complémentaires : d'une part des équipes de recherche « pragmatique » ou « empirique » visent à mettre au point, le plus rapidement possible, des systèmes opérationnels, sans chercher à atteindre une qualité parfaite ;

d'autre part, des équipes développent des idées plus ambitieuses sur le plan théorique, imaginant les prémisses d'une analyse syntaxico-sémantique, mais le plus souvent sans implémentation.

Les équipes dites « empirique » ont, en pratique, surtout conçu des systèmes de traduction mot à mot, avec une analyse linguistique très sommaire. Comme on peut s'en douter, les résultats obtenus étaient très pauvres dans la mesure où ni la syntaxe ni la sémantique n'étaient réellement pris en compte. Dès les années 1950 est apparue la nécessité d'ajouter au moins des informations contextuelles au niveau des dictionnaires, afin de permettre une gestion minimale de l'ambiguïté lexicale. Ces systèmes resteront cependant peu convaincants, ce qui aboutira (si on résume à grands traits) à l'abandon brutal de ce type de recherche dans les années 1960, au moins aux Etats-Unis, suite à différents rapports négatifs sur la question (Bar-Hillel, 1959 ; rapport ALAPC, 1966). Il faut toutefois garder plusieurs éléments en tête pour juger ces systèmes : les recherches sur la formalisation du langage étaient encore balbutiantes, les succès en matière de décryptage (de messages secrets pendant la guerre notamment) et de langages formels (pour programmer les ordinateurs alors naissants) laissaient espérer le même type de succès vis-à-vis des langues humaines : c'est que l'ambiguïté, inhérente aux langues naturelles, avait en grande partie été sous-estimée. Enfin, le rapport ALPAC souligne que les systèmes automatiques sont très loin de la qualité des traducteurs humains qu'ils sont censés remplacer. On peut voir là soit une critique injuste, soit une grave méconnaissance du domaine de la recherche : la traduction automatique était un programme de recherche et il était évident que les systèmes ne pouvaient rivaliser à l'époque avec des traducteurs humains[1].

De ce point de vue, les critiques formulées dans le rapport de Bar-Hillel (1959) sont plus intéressantes. Bar-Hillel remarque d'abord que la traduction nécessite une analyse syntaxique complète du texte, ce qui n'était pas encore complètement évident pour tous les groupes impliqués dans le domaine à l'époque. D'autre part, la traduction nécessite de résoudre les ambiguïtés lexicales (c'est-à-dire essentiellement le choix des mots dans la langue cible), ce qui est au-delà de l'état de l'art et, selon lui, ne semble pas soluble à moyen terme. Une annexe du rapport au titre évocateur (« *A demonstration of the non-feasibility of fully automatic, high quality translation* ») entend montrer que le sens de certains mots ambigus ne peut être déterminé, même en prenant en compte le contexte, ce qui suffit à ruiner l'objectif d'une traduction automatique de qualité. Bar-Hillel prend l'exemple suivant, que nous citons car il demeure un des plus connus de l'histoire de la traduction automatique :

> *« Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy. »*

Pour comprendre la phrase, il faut identifier le fait que « *pen* » désigne ici un parc où l'enfant peut jouer et en aucun cas un stylo. Or, rien dans le contexte ne permet d'inférer automatiquement ce sens pour « *pen* », qui est beaucoup moins fréquent que « stylo ». D'où l'impossibilité pour tous les systèmes de résoudre ce type de cas, qui serait fréquent d'après Bar-Hillel. Il est donc impossible d'envisager une traduction complètement automatique de haute qualité à court ou moyen terme (*FAHQT : Fully Automatic High quality machine translation*).

[1] Il faut toutefois souligner que les chercheurs eux-mêmes avaient promis des résultats beaucoup trop optimistes à très court terme, pour des raisons budgétaires notamment, ce qui a aussi contribué à discréditer le domaine pendant longtemps.

Comme on le sait, ces différents rapports vont avoir pour conséquence un arrêt brutal des financement et donc des recherches en la matière aux Etats-Unis à partir des années 1960. D'autres groupes continueront, notamment en Europe, mais la traduction automatique restera longtemps délaissée, du fait de ces obstacles *a priori* insurmontables. La recherche s'oriente alors davantage vers le traitement automatique des langues, c'est-à-dire vers des recherches plus fondamentales cherchant par exemple à donner une représentation syntaxico-sémantique des phrases. On peut ainsi citer les formalismes à base d'unification qui seront très largement répandus à partir de la fin des années 1970 et qui peuvent servir de base à des systèmes de traduction automatique, comme dans le cas du programme européen Eurotra (Hutchins, 1992).

Les résultats de ces systèmes fondés sur une analyse complète des phrases à traduire seront eux aussi très mitigés. Ils demandent en premier lieu un travail considérable pour décrire les langues visées (dictionnaire de la langue source, de la langue cible, règles d'analyse de la langue source, règles de transfert de langue à langue, etc.). De fait, ces systèmes n'ont quasiment jamais pu être opérationnels à large échelle : la masse de connaissances demandée rend très difficile la couverture au-delà de quelques domaines spécialisés. Plus fondamentalement, l'écriture de règles de désambiguïsation ou de transfert doit tenir compte du contexte. Or, il n'existe pas de définition formelle de la notion de contexte et les différentes réalisations d'un mot ou d'un syntagme en situation de traduction sont très difficiles à prévoir de façon exhaustive pour un humain. On constate que ce type de projet de recherche s'est généralement tourné vers certains phénomènes syntaxiques difficiles, négligeant assez largement l'objectif d'un système de traduction opérationnel à large couverture. A l'opposé, on soulignera toutefois l'expérience de Systran qui, ayant capitalisé des années de recherche et développement, peut dès les années 1980 proposer un système de traduction généraliste de qualité acceptable. En pratique, le système, même avec une bonne couverture de base, demande une constante adaptation au domaine ou au contexte d'utilisation.

2.2 Les systèmes de traduction statistique

La traduction automatique statistique (Koehn, 2009) voit le jour à la fin des années 1980 quand une équipe d'IBM essaie d'appliquer à un problème de traduction des techniques issues de la reconnaissance de la parole. Dans ce dernier domaine, la tâche consiste à produire une séquence de mots à partir d'un signal sonore. La traduction automatique pose un problème en quelque sorte similaire, dans la mesure où il s'agit de produire une séquence de mots dans une langue cible à partir d'une séquence dans une langue source (Brown *et al.*, 1003).

La traduction automatique repose sur la disponibilité de grands corpus bilingues alignés, c'est-à-dire de très grands ensembles de textes en situation de traduction « fidèle » où chaque phrase de la langue source est alignée avec une phrase de la langue cible (plus rarement, on peut avoir affaire à des alignements asymétriques, où une phrase de la langue source correspondant à deux phrases de la langue cible par exemple). L'approche statistique revient à repérer les cooccurrences les plus fréquentes entre langue source et langue cible et à en inférer des traductions potentielles au niveau des mots.

Schématiquement, chaque alignement et chaque correspondance lexicale a au début la même probabilité[2] mais le fait que deux mots apparaissent régulièrement en regard l'un de l'autre va progressivement renforcer leur probabilité d'être traduction l'un de l'autre, ainsi que la probabilité des alignements possibles au niveau des phrases où ces deux mots sont en correspondance. Pour reprendre une illustration de Koehn (2009), on peut imaginer différentes phrases alignées en français et en anglais. Les figures ci-dessous (**Figure 1** à **Figure 4**) illustrent de manière simplifiée mais relativement parlante nous semble-t-il le processus d'alignement au niveau des mots.

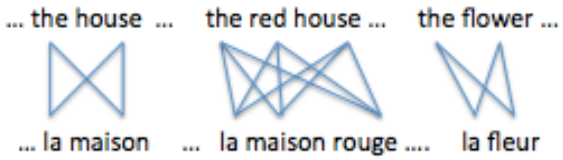

Figure 1 : Initialisation des alignements. Chaque mot anglais est relié à l'ensemble des mots français avec un lien équiprobable.

Figure 2 : Après une première itération, l'algorithme identifie le lien entre « *la* » et « *the* » comme étant le plus probable, sur la base de leur fréquence en langue source et en langue cible (« *la* » est le seul mot qui apparaît systématiquement dans la langue cible quand « *the* » est utilisé dans la langue source). Ces liens sont renforcés (ils apparaissent en gras) au détriments des autres liens et donc aussi des autres alignements possibles.

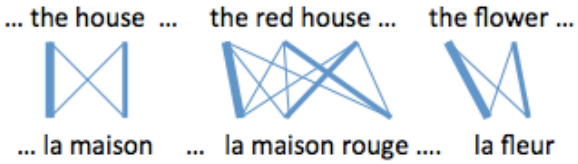

Figure 3 : Après une autre itération, l'algorithme identifie les autres liens les plus probables, entre « *maison* » et « *house* », puis entre « *fleur* » et « *flower* », ainsi qu'entre « *red* » et « *rouge* ». Les autres liens et alignements possibles deviennent de moins en moins probables.

[2] L'équipe d'IBM propose différents modèles reposant sur différentes hypothèses. Ainsi, il est possible de concevoir le même type de modèle en accordant de l'importance à la situation du mot dans la phrase. Intuitivement, cela veut dire qu'un mot situé au début de la phrase source a plus de chance d'être en correspondance avec un mot au début de la phrase cible, et vice versa (Brown *et al.*, 1003). Nous laissons de côté ce type de détails qui n'a pas d'importance ici.

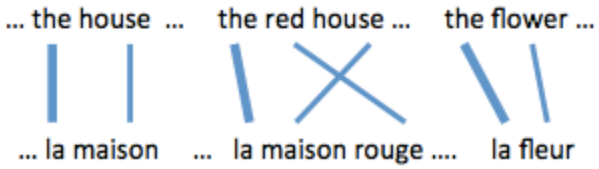

Figure 4 : Le processus se termine quand il y a convergence, c'est-à-dire quand une structure stable a été trouvée. Les autres liens sont ici supprimés mais ils restent en fait disponibles avec une très faible probabilité. Il est possible d'utiliser des techniques de seuil ou de préférence pour ne sélectionner qu'un nombre limité de possibilités, comme sur cette figure où les liens alternatifs ont été complètement effacés.

Comme on le voit aisément sur l'exemple simplifié qui précède, cette stratégie revient à effectuer une traduction mot à mot, suite à la recherche préalable de correspondances lexicales au niveau de très grands corpus. Le modèle peut toutefois être complexifié pour tenir compte de séquences : un des grands enjeux de la recherche dans les années 2000 consistait justement à étendre les modèles IBM originaux pour tenir compte de « fenêtre de traduction » plus longues. Il faut enfin noter que l'exemple ci-dessus semble laisser penser qu'il n'y a qu'une seule traduction possible par mot *in fine*. Ce n'est évidemment pas le cas : l'algorithme permet en fait d'identifier de multiples correspondances pour un mot donné, chacune de ces correspondances ayant une probabilité propre et un contexte d'apparition particulier.

Il n'empêche, cette approche semble à l'opposé des recommandations des rapports des années 1960, à savoir que pour traduire un texte (ou, du moins, une phrase) il faut d'abord le comprendre, c'est-à-dire en donner une représentation syntaxico-sémantique explicite. Les modèles IBM vont cependant connaître un succès phénoménal, au-delà des attentes de l'équipe IBM elle-même, qui ne voyait là initialement que l'occasion d'une expérience simple et basique, mais nécessitant rapidement l'intégration de connaissances linguistiques fines et complexes pour atteindre une qualité acceptable. L'intégration de connaissances extérieures (ressources dictionnairiques notamment) finira par se faire, mais pendant longtemps les principales extensions de ces modèles porteront sur les aspects statistiques (Koehn, 2009).

Pourquoi ces systèmes semblent-ils donner des performances acceptables sur des textes tout venant, en tout cas souvent meilleurs que les systèmes à base de règles mis au point par de nombreux linguistes pendant plusieurs années ? La réponse a en partie été donnée dans la section précédente : il est quasi impossible à des linguistes de définir des règles d'analyse ou de transfert à large couverture, tenant compte du contexte pour un domaine ouvert. A l'inverse, les systèmes statistiques peuvent analyser des corpus de plusieurs millions (voire aujourd'hui milliards) de mots, en tirer des informations dépendant du contexte local (souvent limité à une fenêtre de trois ou quatre mots autour du mot analysé), ce qui suffit à élaborer des systèmes plus robustes et souvent plus exacts que des systèmes développés manuellement.

3 La place de la sémantique

La situation actuelle de la traduction automatique pose plusieurs questions fondamentales pour le domaine. La sémantique est-elle nécessaire pour traduire ou peut-on se contenter de statistiques ? Et d'ailleurs, peut-on dire que les systèmes actuels, notamment ceux reposant sur une base statistique, sont complètement

dépourvus de sémantique ? Enfin, les approches utilisées aujourd'hui vont-elles encore permettre des progrès importants ou bien peut-on au contraire prévoir une limite infranchissable en terme de qualité, du fait des caractéristiques propres des méthodes employées ?

Ces questions sont parfois discutées par les experts du domaine mais elles sont plus souvent reléguées en arrière-plan. Les enjeux commerciaux de la traduction automatique sont importants, comme on le sait. De plus, les approches sont devenues très techniques, à base d'apprentissage artificiel pour l'essentiel, et les préoccupations linguistiques ont été reléguées au second plan. Enfin, ce sont essentiellement les campagnes d'évaluation annuelles qui guident les avancées, ce qui implique un rythme de recherche quasi continu et ne pousse pas à prendre du recul sur le plan épistémologique.

Il est aussi important de rappeler que la période récente a vu des avancées réelles, parfois spectaculaires, et dans tous les cas indiscutables dans le domaine de la traduction automatique. Les méthodes statistiques ont permis de mieux traiter un grand nombre de phénomènes fréquents et importants (recherche du meilleur équivalent traductionnel, gestion des ambiguïtés locales, imbrication de contraintes linguistiques multiples) qui n'avaient pas été bien pris en compte jusque là. Mais le succès des méthodes statistiques a été tel que certains de ses promoteurs n'ont pas hésité à mettre en discussion les progrès récents[3] : certes, les phénomènes linguistiques locaux sont aujourd'hui globalement bien analysés, mais n'a-t-on pas oublié d'autres phénomènes, plus complexe mais néanmoins essentiels, pour permettre de nouvelles avancées ? Certaines limites des systèmes opérationnels disponibles sur le Web par exemple ne sont-elles pas très directement liées aux approches employées ? Mais revenons d'abord aux liens entre sémantique et statistique.

3.1 Les statistiques n'excluent pas la sémantique

Un aspect intéressant des recherches actuelles concerne le statut de l'analyse statistique. On oppose traditionnellement statistique et sémantique : d'un côté le calcul, de l'autre la représentation du contenu des mots ou des phrases. Or, cette opposition est trop simpliste. En effet les statistiques permettent de rendre compte des différents sens des mots en fonction des contextes d'usage. D'ailleurs, qu'est-ce que le sens d'un mot ? Comment le représenter ? On sait en fait très mal définir le sens des mots dans l'absolu (ce qui poussera D. Kayser (1987) à parler d'une « sémantique qui n'a pas de sens ») : les dictionnaires s'y essaient mais le découpage qu'ils proposent est subjectif et ne correspond pas toujours aux emplois en contexte. Par exemple on remarquera que les définitions varient notablement d'un dictionnaire à l'autre, surtout pour les notions abstraites ou les mots fonctionnels.

A l'inverse, n'importe quel locuteur d'une langue saura fournir des synonymes et des exemples d'emploi d'un mot donné. Les différents emplois des mots correspondent en fait à différents usages en contexte : toute la difficulté consiste alors bien évidemment à définir et caractériser la notion de « contexte ». En clair : comment déterminer les différents sens d'un mot donné juste en observant son usage dans un très grand corpus ?

[3] Voir par exemple, l'article de Kenneth Church « *A pendulum swung too far* » (2011), ce que l'on pourrait traduire par « un mouvement de balancier trop brutal ». L'idée principale défendue dans cet article est que l'attrait pour les méthodes statistiques à partir des années 1990 a largement détourné les chercheurs des aspects plus fondamentaux de la langue, exigeant une analyse profonde.

Comment repérer des régularités d'usage ? C'est la tâche du lexicologue qui manipule généralement une multitude d'outils et de critères pour essayer de définir un découpage en sens, qui soit complet et cohérent. Les statistiques permettent de le faire automatiquement, sur une base sans doute très différente de celle du lexicologue, mais avec un résultat souvent pertinent et intéressant.

Les différentes techniques statistiques employées pour aligner les mots et les syntagmes reviennent en fait à calculer des liens de proximité entre mots, c'est-à-dire des équivalences dépendant du contexte. Mettre ainsi en rapport des mots ou des expressions plus complexes entre deux langues, c'est déjà faire de la sémantique. En effet, grâce à l'alignement lexical, plus un mot est ambigu, plus il sera mis en correspondance avec des mots variés dans la langue cible. Plus une expression est figée (comme « *pomme de terre* »), plus elle sera reconnue en tant que telle par les algorithme d'alignement et, le cas échéant, mise en rapport avec un mot simple dans la langue cible (« *potatoe* » par exemple si on est dans un cadre d'alignement français-anglais). L'analyse statistique aboutit donc à une modélisation directe de ces phénomènes de polysémie et de figement, sans théorie linguistique préalable.

On peut aussi admettre que le type de représentation obtenue par analyse statistique est plus adéquate que celle fournie par bien des approches formelles ou théoriques particulières : les notions de sens, de frontière de sens et de figement sont étroitement liées à l'usage et ne sont pas des notions absolues. Ainsi, l'analyse statistique permet de définir différentes granularités de sens (c'est-à-dire définir un nombre plus ou moins grands de sens pour un mot donné, suivant la finesse de l'analyse désirée), ce qui semble bien correspondre à la réalité linguistique (plus que le découpage fixe offert par certains dictionnaires courants en tout cas[4]). Les statistiques rendent compte de façon à la fois simple et subtile de ces phénomènes complexes. Il s'agit là de questions au cœur de la sémantique.

3.2 Les limites des modèles statistiques et la question de la diversité des langues

Nous avons souligné dans la section qui précède le fait que les statistiques n'excluent pas une certaine forme de sémantique. Nous souhaitons en contraste souligner ici deux limites techniques fondamentales : l'alignement de textes fonctionne d'autant mieux qu'il y a une certaine proximité entre langue du texte source et langue du texte cible. On peut donc s'interroger sur les performances possibles pour des langues éloignées comme le finnois, le chinois ou l'arabe vis-à-vis du français ou de l'anglais. Enfin, la traduction statistique suppose fondamentalement la disponibilité de corpus d'entraînement bilingues en nombre très important. Ceci n'est pas sans poser problème dès que l'on quitte le cercle très restreint des langues les mieux représentées sur Internet.

Les langues éloignées de l'anglais

Comme on l'a vu, la plupart des systèmes de traduction automatique reposent aujourd'hui sur une base essentiellement statistique. Il faut en premier lieu souligner le fait que les meilleures performances sont enregistrées quand la langue source ou, mieux, la langue cible, est l'anglais. Il ne s'agit pas uniquement d'une question de masse de données : l'anglais est certes de très loin la langue la mieux pourvue, celle pour laquelle

[4] Il faut toutefois noter que la plupart des grands dictionnaires monolingues ou bilingues ont des entrées hiérarchisées ou arborescentes qui correspondent au fonctionnement supposé de notre mémoire lexicale.

on dispose du plus grand nombre de données, corpus et ressources, mais la réussite de la traduction statistique pour l'anglais est aussi due aux qualités propres de cette langue. La morphologie très pauvre et le caractère relativement fixe des mots en anglais en font un candidat idéal pour la tâche, dans la mesure où l'analyse morphologique, la décomposition des formes lexicales (voire, jusqu'à un certain point, la syntaxe, pour des raisons complémentaires) peuvent être quasiment omises, sans grand dommage.

En outre, le repérage d'équivalents traductionnels au niveau de mots ou de segments fonctionne d'autant mieux qu'on a affaire à des langues proches, qui autorisent un découpage en mots relativement similaire entre langues source et langue cible. Ceci se traduit très directement dans les performances des différents systèmes : ainsi la traduction de l'allemand vers l'anglais fonctionne mieux que la traduction de l'anglais vers l'allemand dans la mesure où les mots composés allemands (qui combinent plusieurs mots simples et un ensemble complexe d'un seul tenant) restent problématiques pour le traitement automatique. Si l'on sait actuellement relativement bien analyser les composés existants, il est beaucoup plus difficile de « générer » des mots composés corrects en allemand, ce qui explique le différentiel de performances suivant le sens de traduction (comme on sait relativement bien analyser les composés allemand, la traduction vers l'anglais n'est pas trop problématique, mais comme on ne sait pas bien générer ces mots composés en allemand, la traduction vers l'allemand reste médiocre).

La traduction vers le japonais ou, plus récemment, le chinois ou l'arabe a suscité un nombre très important de recherches. Les performances, comparées à celles obtenues avec des langues indo-européennes comme le français ou l'espagnol restent moindres, dans la mesure où ces langues ont une structure très éloignée de l'anglais. Pour ces langues, la mise au point de systèmes hybrides intégrant une composante statistique mais aussi des connaissances de nature linguistique poussée sera probablement la principale source de progrès dans les années à venir. Plus généralement, la traduction des langues ayant une morphologie complexe, au premier rang desquelles les langues agglutinantes, reste très médiocre. Il semble difficile de se passer dans ce cas de systèmes d'analyse précis, décomposant les mots pour procéder à l'analyse de leur nature et de leur fonction dans la phrase. C'est en ce sens qu'on a parlé de l'anglais comme d'un candidat idéal pour la tâche au début de cette section.

Le cas des langues rares, ou le retour de la langue pivot

Sur un autre plan, il faut noter que pour pouvoir être mis au point, tous les systèmes statistiques nécessitent de gigantesques masses de textes bilingues. Ces approches fonctionnent d'autant mieux que la masse de texte à disposition est grande (des corpus de plusieurs millions de mots sont nécessaires pour obtenir un résultat de qualité correcte). A ce propos, Mercer, un des membres de l'équipe IBM à l'origine de la traduction statistique, avait ainsi proclamé : « *there is no data like more data* », ce que B. Habert avait naguère transposé en français par « *gros, c'est beau* » (Habert, 2000). Pour Mercer comme pour tous les tenants de l'approche statistique, la meilleure stratégie pour développer un système consiste à amasser le plus de données possible. Ces données doivent idéalement être représentatives et diversifiées, mais comme il s'agit de critères qualitatifs difficiles à évaluer, c'est le critère quantitatif qui continue de prévaloir. De fait, il a été démontré que les performances des systèmes s'amélioraient régulièrement en fonction de la masse de quantité disponible pour les mettre au point.

Dès lors, il est clair qu'au-delà d'une dizaine ou une quinzaine de langues bien répandues sur Internet, les performances des systèmes baissent considérablement, surtout si une des langues (source ou cible) n'est pas l'anglais. La masse de données est alors tout simplement insuffisante pour obtenir un système performant. Des techniques essaient de pallier le manque de données bilingues en prenant mieux en compte des connaissances obtenues à partir de grand corpus monolingues mais ceci reste insuffisant pour obtenir une traduction de qualité. Une autre stratégie, très populaire, consiste à essayer de concevoir des systèmes de traduction passant par l'anglais comme langue pivot, pour concevoir des traductions de « langue rare » (ou « peu dotée » sur le plan des ressources électroniques) à « langue rare ». L'intuition est que s'il n'existe pas assez de données bilingues entre deux langues (par exemple, entre le grec et le finnois), la difficulté peut être en partie contournée en procédant par une traduction du grec vers l'anglais puis de l'anglais vers le finnois. Cette stratégie est relativement simple à mettre en œuvre sur le plan technique mais ne fait que « contourner » le problème : la traduction de et vers l'anglais n'exclut pas les erreurs et la mise en œuvre de deux étapes de traduction au lieu d'une multiplie aussi les erreurs[5].

Beaucoup de commentateurs ont ainsi remarqué que Google Traduction faisait de plus en plus usage de l'anglais comme langue pivot. Ceci aboutit à des résultats approximatifs : comme l'a relevé par exemple Frédéric Kaplan sur son blog[6], avec Google Traduction, « *Il pleut des cordes* » se transforme en « *Piove cani and gatti* » en italien. De même, « *Cette fille est jolie* » devient « *Questa ragazza è abbastanza* » (« *cette fille est moyenne* »). Ces erreurs grossières sont dues au passage par l'anglais : pour « *Il pleut des cordes* », Google identifie l'équivalent anglais « *It rains cats and dogs* » et pour « *joli* », il faut identifier « *pretty* » qui peut correspondre en italien à « *abbastanza* » mais qui, pour le coup, n'est plus un équivalent de « *joli* » ! Ces exemples perdent rapidement de leur actualité car Google corrige en permanence son système : l'article de F. Kaplan date du 15 novembre 2014, et au 1er décembre, la traduction de « *Il pleut des cordes* » en italien était devenue « *Piove a dirotto* ». Plus fondamentalement, Google travaille en permanence à l'amélioration des équivalences traductionnelles au niveau des expressions complexes (comme dans le cas de « *It rains cats and dogs* » par exemple). Certains éléments sont de plus faciles à repérer dans les ressources linguistiques, quand une expression idiomatique d'un côté n'a pas d'équivalent de l'autre (« *It rains cats and dogs* » est repéré comme expression idiomatique en anglais, tout comme « *pleuvoir des cordes* ») en français, mais est visiblement traduit de façon littérale en italien). Les pivots ambigus (cas de « *pretty* ») posent un problème plus compliqué : ils sont faciles à repérer mais difficiles à corriger car l'ambiguïté est inhérente à la langue.

Nous avons pris ici l'exemple d'un des systèmes de traduction les plus populaires sur la Toile : notre but n'est pas de mettre en évidence quelques erreurs anecdotiques ou de collectionner des remarques éparses. Ces problèmes montrent en fait que des questions classiques de sémantique lexicale continuent bien évidemment de se poser, en dépit des progrès récents.

---

[5] On sait que la traduction d'un texte de l'anglais vers le français puis du français vers l'anglais, relancée plusieurs fois, a tendance à augmenter les erreurs et à s'écarter progressivement du texte original, jusqu'à aboutir parfois à un résultat incompréhensible. Ce type de problème est évidemment augmenté quand une des langues est moins représentée sur Internet, comme dans notre exemple entre le grec et le finnois.

[6] https://fkaplan.wordpress.com/2014/11/15/langlais-comme-langue-pivot-ou-limperialisme-linguistique-cache-de-google-translate/

4 Conclusion : une sémantique de l'usage ?

Nous avons essayé dans cet article de fournir quelques pistes permettant de mieux comprendre comment des systèmes de traduction pouvaient fonctionner sans représentation explicite du sens. En fait, cette formulation est relativement inadéquate : les systèmes de traduction purement statistiques n'ont pas recours à des ressources prédéfinies proposant un découpage *a priori* du sens des mots, mais ces systèmes, par l'analyse de très grands corpus, permettent le calcul de contextes d'usage plus ou moins fins, ce qui revient à calculer des cas d'usage à granularité variable.

Cette approche implémente donc de façon très directe la célèbre intuition de Firth, éternellement reprise dans le domaine de la linguistique de corpus : « on connaît un mot par son contexte » (« *You shall know a word by the company it keeps* », Firth 1957:11). C'est évidemment la notion de contexte qui reste difficile à cerner : élargir la fenêtre de calcul (c'est-à-dire d'un nombre de mot plus ou moins grand autour du mot étudié) améliore les résultats mais est très coûteux sur le plan computationnel. L'intégration d'autres informations (dont des ressources extérieures comme des dictionnaires et des grammaires) peut contribuer à améliorer la qualité des résultats, ce qui aboutit à des systèmes hybrides comme on l'a déjà souligné.

Au-delà, l'approche purement statistique nous semble intéressante dans la mesure où elle souligne la force de certains éléments connus mais trop souvent sous-estimés, comme la nature fondamentalement statistique (voire « zipfienne ») des langues ou le caractère continu de la sémantique. A l'inverse, cette approche semble aussi montrer l'inadéquation des modes de représentation logiques, souvent trop rigides face à la plasticité de la langue. L'ambiguïté n'est pas en soi un problème qui devrait être corrigé (ou qui ne peut être résolu qu'à travers l'élaboration de listes infinies), c'est en fait le reflet de différents usages des mots dans différents contextes. Chaque contexte est particulier mais il est possible d'identifier des cas d'usage réguliers par généralisation, ce que les statistiques font très bien (Poibeau, 2014). Il nous semble qu'il y a toutefois au moins deux limites à cette approche statistique : d'une part, comme nous l'avons déjà souligné, le fait que le contexte pris en compte par les algorithmes actuels reste limité ; définir le contexte idéal, sa nature et son contenu, reste un problème entier. D'autre part, le langage est affaire d'usage et par définition, le texte est un objet figé qui exclut tout le contexte communicationnel extralinguistique. Il est douteux que l'ensemble des connaissances nécessaires au traitement automatique de la langue puisse être inféré automatiquement et directement à partir de corpus, même si ceux-ci se composent de milliards de mots. C'est d'ailleurs en partie pour répondre à ce problème que les systèmes statistiques cherchent de plus en plus à intégrer des ressources linguistiques prédéfinies, même s'il faut ensuite s'assurer de la cohérence de ces données avec celles des modèles statistiques. Il y a là un champ de recherche à part entière.

Pour conclure, il faut enfin souligner que l'on n'a abordé ici que des problèmes de sémantique lexicale. Peu de systèmes actuels intègrent une analyse de la structure des phrases ou des liens entre phrases, ce qui est une cause majeure d'erreurs en traduction automatique (au sens où les résultats obtenus peuvent être assimilés à des contresens voire, le plus souvent, être complètement incompréhensibles). Ces questions font partie de celles qui ont été largement laissées de côté en traduction automatique, comme le déplore Church dans son article de 2011 déjà cité : elles posent pourtant des problèmes aussi fondamentaux que la sémantique lexicale et sont probablement plus difficiles

encore à résoudre. C'est en ce sens que Church dit que la traduction automatique, et plus généralement les approches statistiques en matière de traitement automatique des langues, ont jusqu'ici ramassé les « fruits les plus faciles à atteindre » (l'expression « *low hanging fruits* » est devenue très populaire dans le domaine), c'est-à-dire qu'elles se sont intéressées aux problèmes les plus abordables. Il existe cependant de nombreux fruits plus difficiles à atteindre, qu'il faudra pourtant aborder quand les tâches les plus faciles auront été résolues de manière raisonnable. Il n'est alors pas évident que les statistiques suffiront.

Références

BAR-HILLEL J. (1959), « Report on the state of machine translation in the United States and Great Britain », Rapport technique pour le US Office of Naval research Information. Disponible en ligne : http://www.mt-archive.info/Bar-Hillel-1959.pdf.

BROWN P. DELLA PIETRA V., DELLA PIETRA S. et MERCER R. (1993), « The mathematics of statistical machine translation: parameter estimation », *Computational Linguistics*, Vol. 19, n°2, pp. 263-311.

CHURCH K. (2011), « A pendulum swung too far », *Linguistic Issues in Language Technology*, Vol. 6, n°5.

COMITÉ ALPAC (1966), *Language and Machines — Computers in Translation and Linguistics. ALPAC report*. Washington DC: National Academy of Sciences & National Research Council.

FIRTH J. (1957), *Papers in Linguistics (1934-1951)*, Oxford: Oxford University Press.

HABERT B. (2000), « Des corpus représentatifs : de quoi, pour quoi, comment ? ». In BILGER, M. (éd.), *Linguistique sur corpus - Etudes et réflexions*, Perpignan, Perpignan : Presses Universitaires.

HUTCHINS J. (1986), *Machine translation: past, present, future*, Chichester: Ellis Horwood (Ellis Horwood Series in Computers and their Applications).

HUTCHINS J. (1992), « Eurotra ». In Hutchins and SOMMERS (1992), pp. 239-258.

HUTCHINS J. (2003), « ALPAC, the (in)famous report », In NIRENBURG *et al.* (2003), pp. 131-135.

HUTCHINS J. et SOMERS H. (1992), *An introduction to machine translation.* Londres : Academic Press.

KAYSER D. (1987), « Une sémantique qui n'a pas de sens », *Langages*, n° 87, pp. 33–45.

KOEHN P. (2009), *Statistical Machine Translation*, Cambridge: Cambridge University Press.

NIRENBURG S., SOMERS H. et WILKS Y., éditeurs (2003), *Readings in Machine Translation*, Cambridge, USA : MIT Press.

POIBEAU T. (2014), « La linguistique est-elle soluble dans la statistique ? », *Revue Sciences et Lettres*. Vol. 2. Disponible en ligne : http://rsl.revues.org/402.

Titre en anglais : Translating without understanding: What kind of semantics for machine translation?

Summary : The translation activity involves understanding the text to be translated so as to transpose the main ideas as precisely as possible in the target language. It is largely assumed that the first generations of

machine translation systems (1950-1965) failed because of the absence of semantic analysis or at least because of weaknesses in their semantic analysis component. Research has then largely focused on text understanding, in order to be able to calculate a relevant semantic representation of the text. Contrary to this approach, the late 1980s and 1990s have seen new kinds of systems based on a purely statistical analysis, with no explicit semantic representation of the textual content. In this article we will investigate the attitude of current systems towards semantics. To what extent do current systems based on large collections of texts (the « big data » approach) integrate semantic information?

Keywords: machine translation; semantics; evaluation