Revue de Traduction et Langues Volume21 Numéro1/2022
Journal of Translation Languages مجلة الترجمة واللغات
ISSN (Print): 1112-3974 EISSN (Online): 2600-6235

Traitement de l'ambiguïté syntaxique et sémantique en TA neuronale : analyse de la traduction de l'anglais vers le français, l'espagnol et l'italien

Syntactic and Semantic Ambiguity Processing in Neural MT from English to French, Spanish and Italian

François Maniez
Université Lumière Lyon 2- France
francois.maniez@univ-lyon2.fr
Centre de recherche en linguistique appliquée-CeRLA
0000-0003-1704-1493

Comment citer cet article:
Maniez, F. (2022). Traitement de l'ambiguïté syntaxique et sémantique en TA neuronale : analyse de la traduction de l'anglais vers le français, l'espagnol et l'italien. *Revue Traduction et Langues21* (1), 10-27.

Reçu : 22/05/2022 ; **Accepté :** 21/08/2022, **Publié :** 31/08/2022

L'auteur correspondant : François Maniez

Keywords

English, machine translation, noun phrase, Romance languages, semantic ambiguity, syntactic ambiguity.

Abstract

Despite recent advances in artificial intelligence, human translators outperform MT for at least three types of tasks: identifying referents in anaphora (especially of the interphrastic kind), resolving semantic ambiguity (which is mainly due to polysemy or homonymy), and resolving syntactic ambiguity (especially with poorly inflected source languages such as English).

Using the results obtained by two freely available online machine translation programs, Google Translate and DeepL, we examine how these two types of ambiguity are processed in translation from English into French, Spanish and Italian.

Our results show that the two programs perform well overall in resolving the simplest cases of syntactic ambiguity, with difficulties arising more frequently for noun phrases featuring atypical syntactic divisions and rarely used collocations. MT output for ambiguous structures involving verb roots followed by the –ING morpheme (flying planes, growing pains) is studied, as well as syntactic structures in which two or more nouns are preceded by one or more adjectives. MT handles relatively well the longest of those structures (ADJ ADJ N N N N), probably because their subsets are part of the bilingual or target language monolingual corpora that underlie MT systems.

Structures involving head modification and coordination (ADJ N AND N) are also known to pose problems for MT and human translators alike. But since many of the most frequent N AND N structures involve cohyponyms (men and women, brothers and sisters), antonyms (rights and duties, costs and benefits) or near-synonyms (aid and advice), their translation as a whole unit generally triggers the choice of correct syntact dependencies in translation. Structures in which the adjective only modifies the first noun (fresh air and exercise, social sciences and humanities) are much less frequent and are also probably translated as a whole unit. Structures involving premodification, coordination and post-modification may give rise to four distinct types of structures depending on whether long-range dependencies apply (detailed [knowledge and understanding] of the IT industry, [ethnic group] and [place of birth], invaluable [context and [source of information], [close friend] and confidant] of Mr Jones. Structures in which both long-range dependencies apply (integrated prevention and control of pollution) are the ones which most frequently cause errors for MT.

Semantic ambiguity cases have been processed with increasing success by MT, especially when collocates vary widely for the main two meanings of homonyms (a well-known example is the word pen). Processing polysemy (for instance the medical use of conditions in pre-existing conditions) is a bit more of a challenge for MT. Other cases involving concentration of several polysemic terms in the same sentence (Changing the placement of beams relative to the staff involves changing the direction of the stems in the beam) also create difficulties for MT when the polysemic terms are used without any of their usual collocates (here in the specialised field of musical edition).

Homonymy cases involving grammatical category changes (N-to-V or V-to-N conversion) seem to continue to pose the most difficulties to neural MT, despite increasing consideration of intra- and extraphrastic context. Potentially ambiguous word sequences (treatment increase in as the daily dose and duration

of treatment increase), which were processed incorrectly before neuronal MT, are now correctly translated. But word sequences in which one word belongs to a part of speech which is not the most commonly used one (e.g. the noun remains in what remains can be considered) may cause occasional errors. Several examples that involve the verb founder are studied, and they frequently trigger translation of the noun in all three Romance languages (or translations of the verbs find or found due to incorrect segmentation).

Mots clés

Ambiguïté sémantique, ambiguïté syntaxique, anglais, groupe nominal, langues romanes, traduction automatique.

Résumé

Malgré les progrès récents de l'intelligence artificielle, les traducteurs humains obtiennent de meilleurs résultats que la TA pour au moins trois types de tâches : l'identification des référents dans les cas d'anaphore (notamment interphrastique), la résolution de l'ambiguïté sémantique (principalement due à la polysémie ou à l'homonymie) et celle de l'ambiguïté syntaxique (notamment dans le cas de langues sources à flexion limitée comme l'anglais). En nous appuyant sur les résultats obtenus par deux programmes de traduction automatique gratuitement accessibles en ligne, Google Translate et DeepL, nous examinons la façon dont sont traités ces deux types d'ambiguïté dans la traduction de l'anglais vers le français, l'espagnol et l'italien. Nos résultats font apparaître une bonne performance globale des deux logiciels utilisés pour la résolution des cas d'ambiguïté syntaxique les plus simples, les difficultés se présentant plus fréquemment dans le cas de groupes nominaux aux découpages syntaxiques atypiques et contenant des collocations peu usitées. Les cas d'homonymie impliquant un changement de catégorie grammaticale (conversion) semblent être ceux qui continuent de poser le plus de difficultés à la TA neuronale, en dépit d'une prise en compte croissante du contexte intra- et extraphrastique.

1. Introduction

On observe depuis deux décennies un fort progrès qualitatif de la traduction automatique (désormais TA), qui est devenue accessible au grand public suite à la mise en ligne de deux programmes de TA, Google Translate (désormais GT) en 2006 et DeepL (désormais DL) en 2017. Par ailleurs, l'utilisation depuis 2017 de la TA neuronale a considérablement amélioré la fluidité des traductions de l'anglais vers les langues romanes, qui laissait à désirer quand les anciens programmes à base de règles étaient encore appliqués.

Dès le début du siècle, la TA obtenait des résultats honnêtes pour le traitement des textes réglementaires ou scientifiques, car ils présentaient des caractéristiques favorables à un traitement par la TA, à savoir un vocabulaire moins polysémique que celui de la langue dite générale (ce qui engendrait une moindre quantité d'ambiguïtés sémantiques) et constitué pour une partie importante de termes complexes, dont la traduction est généralement fiable et homogène si leurs équivalents de traduction (désormais ET) sont stockés dans les mémoires de traduction qu'utilisent les programmes de TA. Dans la cadre

de la traduction professionnelle, il est de plus en plus fréquent que la traduction de ce type de texte soit effectuée par un programme de TA avant d'être révisée par un humain, cette activité de révision étant généralement désignée par le terme de post-édition, un emprunt adapté de l'anglais post-editing, qui signifie « correction a posteriori » (Robert 2010).

Les traducteurs humains continuent néanmoins d'obtenir de meilleurs résultats que la TA dans plusieurs cas de figure, dont l'identification des référents dans les cas d'anaphore, ainsi que la résolution de l'ambiguïté sémantique et syntaxique. Dans le cas de la traduction de l'anglais vers les langues romanes, l'ambiguïté syntaxique est fréquente, en raison du faible degré de flexion de l'anglais. En effet, les suffixes des verbes réguliers sont de nombre limité (-ed, -ing et –s), et peuvent tous engendrer divers types d'ambiguïté : les formes en -ed peuvent correspondre à l'emploi du simple past ou du participe passé (adjectivé ou non), les formes en –ing peuvent être verbales, nominales ou adjectivales, et dans les cas de conversion verbe□nom ou nom□verbe, les formes en –s peuvent relever de l'une ou de l'autre catégorie.

En ce qui concerne l'anaphore, jusqu'à une période récente, les modèles de la TA traitaient généralement les phrases isolément et ne tenaient compte d'aucune information contextuelle au-delà des limites de la phrase (Tiedemann et Scherrer 2017). Cependant, la TA neuronale a récemment commencé à prendre en compte le contexte interphrastique (Voita et al 2019).

Dans les écrits scientifiques, l'identification des référents pose peu de problèmes pour la TA, en partie à cause de l'absence de contraintes stylistiques qui empêchent la répétition (limitant ainsi l'utilisation de l'anaphore). En revanche, dans la prose journalistique, l'anaphore et l'ellipse sont plus fréquemment utilisées, ce qui peut entraîner un plus grand nombre d'erreurs de la TA.

L'exemple ci-dessous démontre l'absence de cohérence interphrastique (traduction GT, 17/10/2021) générée par l'absence d'identification du référent anaphorique du pronom it (The academy) dans la deuxième phrase, le bigramme it took étant identifié dans sa globalité comme signifiant il a fallu :

> The academy accepted $525,000 in donations from Coke in 2012. The following year it took a $350,000 donation from the company.
> L'Académie a accepté 525 000 $ de dons de Coke en 2012. L'année suivante, il a fallu un don de 350 000 $ de la société.

Pour ce même exemple, la cohérence interphrastique est préservée par DL :
L'Académie a accepté 525 000 $ de dons de Coke en 2012. L'année suivante, elle a accepté un don de 350 000 $ de la société.

Même si la traduction des formes verbales accepted et took par la même forme française (a accepté) provoque une répétition stylistiquement dommageable, cette répétition n'a pas de répercussions sur la cohérence interphrastique, alors que c'est le cas quand l'inverse (traduction du même mot par deux synonymes) se produit, comme dans

Disponible en ligne à https://www.asjp.cerist.dz/en/Articles/155

l'exemple suivant pour lke mot penalty, rendu successivement par amende et sanction (traduction GT, 23/10/2021) :

> The individual mandate is the law's controversial requirement that all Americans maintain qualifying health insurance coverage or pay a penalty. In 2012, the Supreme Court upheld this penalty as an exercise of Congress's taxing power. Le mandat individuel est l'obligation controversée de la loi selon laquelle tous les Américains doivent conserver une couverture d'assurance maladie ou payer une amende. En 2012, la Cour suprême a confirmé cette sanction en tant qu'exercice du pouvoir de taxation du Congrès.
> De nouveau, la traduction de DL préserve davantage la cohérence interphrastique en répétant l'emploi du mot pénalité.

2. Traitement de l'ambiguïté syntaxique

Nous aborderons tout d'abord les structures dans lesquelles une forme en –ING précède un nom, avant de nous pencher sur celles qui impliquent la prémodification, ainsi que sa combinaison avec la coordination et la post-modification.

2.1 La structure V+ING N(s)

De nombreux mots suffixés en *–ing* peuvent appartenir à plusieurs catégories grammaticales, comme c'est le cas dans cet exemple rendu célèbre par Chomsky dans son ouvrage *Syntactic Structures* (1957) :

GT	DL
Les avions volants peuvent être dangereux.	Voler en avion peut être dangereux.
Volar aviones puede ser peligroso.	Volar en avión puede ser peligroso.
Gli aerei in volo possono essere pericolosi.	Far volare gli aerei può essere pericoloso.

Flying planes can be dangerous.

L'ambiguïté de cette phrase disparaît si l'on en supprime le modal CAN, ce qui donne respectivement *Flying planes is/are dangerous*, la première structure impliquant la présence du verbe transitif *fly*, qui signifie ici « piloter ». Le tableau ci-dessous montre la TA de la phrase originale en français, espagnol et italien (traduction GT et DL, 23/10/2021) :

Cet exemple permet de constater que GT n'utilise pas le même découpage syntaxique pour les trois langues : seul l'espagnol traduit *flying* par le verbe transitif *volar*. Pour DL, l'italien est correct, mais les versions française et espagnole sont inexactes et

traduisent plutôt *Flying* ***in*** *a plane can be dangerous*. Cet exemple met également en évidence l'une des faiblesses de la TA, à savoir le recours par défaut à l'ET le plus fréquent (*voler*, *volar*, *volare*) au détriment d'autres mots qui conviendraient mieux dans ce contexte (*piloter*, *pilot(e)ar* ou *pilotare*). Si l'on désambiguïse l'énoncé en employant le singulier (*Flying a plane can be dangerous*), on observe pour les deux logiciels le même type d'imprécision, avc l'emploi de *voler* en français (DL) et de *volare* en italien (GT). Cependant, *I can fly a plane* est correctement traduit dans les six cas de figure.

L'ambiguïté de l'exemple précédent reposait sur la transitivité éventuelle du verbe *fly*. Dans l'exemple de la phrase *The growing pains of the health care act are frustrating patients*, c'est une ambiguïté catégorielle (nom/adjectif) qui constitue un écueil pour la TA (traduction GT et DL, 23/10/2021) :

GT	DL
Les **douleurs croissantes** de la loi sur les soins de santé frustrent les patients.	Les **difficultés de croissance** de la loi sur les soins de santé frustrent les patients.
Los **crecientes dolores** de la ley de asistencia sanitaria frustran a los pacientes.	Los **dolores crecientes** de la ley de asistencia sanitaria están frustrando a los pacientes.
I **dolori crescenti** della legge sull'assistenza sanitaria stanno frustrando i pazienti.	I **dolori di crescita** della legge sull'assistenza sanitaria stanno frustrando i pazienti.

Grow est ici un verbe intransitif, si bien que la structure syntaxique de cette phrase n'est pas ambiguë à cet égard. Ici, *growing* est en fait le nom *croissance*, si bien que la traduction proposée par DL en français et en italien a le mérite de proposer à la post-édition une interprétation syntaxiquement correcte, contrairement aux quatre autres, dans lesquelles *growing* est traduit par un adjectif.

2.2 Structures utilisant la prémodification

La prémodification permet une éventuelle modification à distance d'un nom par un adjectif (ou un nom) antéposé. De ce fait, la structure ADJ – N – N est susceptible de deux découpages, selon que l'adjectif modifie uniquement le nom qui le suit comme dans le groupe nominal *[social security] benefits*, ou l'ensemble constitué par les deux noms, comme dans *unfair [contract terms]*. Dans la mesure où les bigrammes et les trigrammes impliqués font généralement partie des mémoires de traduction utilisées par la TA, ces structures ne posent pas de problème aux deux logiciels.

Dans la prose réglementaire ou scientifique, l'ajout successif d'élements prémodificateurs mène à d'autres structures de complexité croissante. Ainsi, l'ajout d'un adjectif supplémentaire (ADJ ADJ N N) crée un autre découpage possible, soit trois au total. L'ajout d'un nom à droite de cette dernière structure (ADJ ADJ N N N) donne six

découpages possibles, et l'on arrive même à dix découpages théoriquement possibles avec l'ajout d'un autre nom (ADJ ADJ N N N N).
Cependant, même quand la liste des éléments prémodificateurs s'allonge, les résultats obtenus par la TA restent corrects dans la mesure où les bigrammes et les trigrammes qui composent les termes complexes sont correctement interprétés. Ainsi, la prose scientifique anglaise (notamment en langue médicale) contient de très nombreux exemples de GN de structure de longueur 5 ou 6, qui sont pour la plupart correctement traités par la TA. L'exemple ci-dessous donne le résultat de la traduction d'un GN de structure ADJ ADJ N N N N dans la phrase ***Significant synergistic tumor growth inhibition effect*** *was demonstrated […]* (traduction GT et DL, 23/10/2021) :

GT	DL
Un **effet synergique significatif d'inhibition de la croissance tumorale** a été démontré […].	Un **effet synergique significatif d'inhibition de la croissance tumorale** a été démontré […].
Se demostró un **efecto de inhibición sinérgico significativo del crecimiento tumoral** […].	Se demostró un **efecto sinérgico significativo de inhibición del crecimiento tumoral** [...].
Un **significativo effetto sinergico di inibizione della crescita tumorale** è stato dimostrato […].	Un **significativo effetto sinergico di inibizione della crescita tumorale** è stato dimostrato [...].

On remarque ici une forte similitude entre GT et DL, jusque dans l'utilisation de l'adjectif relationnel *tumoral(e)* pour traduire le nom *tumor*. En espagnol, seul l'ordre des deux adjectifs prémodificateurs varie, et on observe en italien la même similitude qu'en français entre les deux traductions employées.

2.3 Structures combinant la prémodification et la coordination

L'ajout d'une conjonction de coordination à la première structure précédemment mentionnée pour former un groupe nominal de structure *ADJ N and N* aboutit également à deux découpages possibles. Ainsi, dans le GN *old friends and neighbors*, l'adjectif *old* peut modifier uniquement le premier nom ou l'ensemble des deux noms coordonnés, ce qui se produit dans la majorité des exemples de cette structure.

Dans le *British National Corpus*, les deux noms des structures qui sont le plus fréquemment employées sont des co-hyponymes (*men and women, brothers and sisters*), des antonymes (*rights and duties*, *costs and benefits*) ou des mots de sens proche (*aid and advice*). La structure dans laquelle l'adjectif ne modifie que le premier nom ne concerne qu'environ 15 % des cas des groupes nominaux employés au moins dix fois dans le corpus (les exemples les plus fréquents du corpus comprennent des GN tels que *fresh air and exercise* ou *social sciences and humanities*).

Disponible en ligne à https://www.asjp.cerist.dz/en/Articles/155

Dans la mesure où là encore, les quadrigrammes impliqués font généralement partie des mémoires de traduction utilisées par la TA, ces structures sont généralement bien traitées par les deux logiciels.

2.4 Structures combinant la prémodification, la coordination et la post-modification

Outre la coordination, certaines structures encore plus complexes font également intervenir la post-modification (*ADJ N and N of N*). On obtient alors quatre découpages syntaxiques distincts :

Type	Exemple
A	detailed [knowledge and understanding] of the IT industry
B	[ethnic group] and [place of birth]
C	invaluable [context and [source of information]]
D	[[close friend] and confidant] of Mr Jones

Pour cette structure, les quatre découpages syntaxiques sont les suivants : le type A dans lequel les deux dépendances à distance s'appliquent, le type B dans lequel aucune ne s'applique, et les types C et D, dans lesquels l'une s'applique et l'autre non.

Lors d'une étude précédemment effectuée sur la traduction par GT de phrases contenant cette structure (Maniez & Villar-Diaz 2018), la majorité des erreurs observées concernaient des structures de type A dans lesquelles une seule des deux dépendances syntaxiques à distance était correctement traduite, l'erreur la plus fréquente consistant à occulter la prémodification du deuxième nom par l'adjectif (traduction par une structure de type D), comme dans l'exemple suivant (traduction 05/02/2018) :

The integrated prevention and control of pollution, **which is the aim of this directive, […]**	**La prévention intégrée et le contrôle de la pollution, qui est l'objectif de cette directive, […]**	**La prevención integrada y el control de la contaminación, que es el objetivo de esta directiva, […]**

Les traductions plus récentes, elles, tiennent bien compte de la dépendance à distance, et ce dans les trois langues de notre étude.

L'absence de prise en compte de cette dépendance à distance peut être due à la prépondérance statistique du bigramme constituant l'ET le plus fréquent pour traduire l'ensemble constitué par l'adjectif et le premier nom de ces groupes nominaux. En effet, il est probable que lorsque ce bigramme est d'usage très fréquent, il sera sélectionné comme l'équivalent de traduction le plus probable. Ainsi, dans l'expression *mutual [recognition and enforcement] of judgments*, la fréquence très élevée dans les corpus français du bigramme *reconnaissance mutuelle* par rapport à celle du bigramme *exécution mutuelle* (traduction possible de *mutual enforcement*) pourrait être à la source de l'erreur

observée pour les deux programmes de TA dans les trois langues, comme c'est le cas dans l'exemple ci-dessous :

The principle of mutual recognition and enforcement of judgments is the cornerstone of judicial and police cooperation.	**Le principe de reconnaissance mutuelle et d'exécution des jugements est la pierre angulaire de la coopération judiciaire et policière.**	**El principio de reconocimiento mutuo y ejecución de las sentencias es la piedra angular de la cooperación judicial y policial.**	**Il principio del riconoscimento reciproco e dell'esecuzione delle decisioni è la pietra angolare della cooperazione giudiziaria e di polizia.**

Les structures de type C donnent également souvent lieu à des erreurs, qui sont manifestes quand les deux noms coordonnés sont de genre grammatical distinct. Ainsi, aucun des résultats donnés par la TA pour la traduction du groupe nominal *a strong [personality and [sense of identity]]* n'est réellement satisfaisant (traduction GT et DL, 25/10/2021) :

The peoples of Europe who are identifiable as stateless nations, or in regions with a strong personality and sense of identity, […]	**Les peuples d'Europe identifiables comme nations apatrides, ou dans des régions à forte personnalité et sentiment d'identité, [...]**	**Les peuples d'Europe qui sont identifiables en tant que nations sans État, ou dans des régions ayant une forte personnalité et un sentiment d'identité, [...].**

Les traductions espagnoles et italiennes du groupe prépositionnel final (*en regiones con una fuerte personalidad y sentido de identidad* / *in regioni con una forte personalità e senso di identità*) reproduisent cette absence de prise en compte de la dépendance syntaxique à distance.

3. Traitement de l'ambiguïté sémantique

La désambiguïsation sémantique est un enjeu important de la TA depuis plusieurs décennies. Les premières méthodes de désambiguïsation sémantique automatique se fondaient sur la présence dans le contexte de mots appartenant au même champ sémantique que les polysèmes, en se fondant sur leur présence dans la définition du poysème dans les dictionnaires électroniques (Kilgarriff & Rosenzweig 2000, Michiels 1999, Brun et al. 2005). Pour prendre un exemple souvent cité, celui du mot pen, on utilisait comme indice la présence dans le contexte de mots figurant dans les définitions de ses deux sens (par exemple, ink et paper pour le sens « stylo » ou pigs et cows pour le sens « enclos »).

3.1 L'homonymie

Les problèmes que pose l'homonymie ont donc plus de chances d'être résolus par la TA dans des phrases longues dont le contexte éclaircit le sens des polysèmes, comme il le fait pour les humains. Ainsi, la phrase Take it to the pens, que l'on peut traduire par Emmenez-le/la aux enclos (en parlant d'un animal), était traduite en 2018 par Prenez-le aux stylos (GT) et Emmenez-le aux stylos (GT). Les résultats obtenus très récemment montrent que le sens « stylo » est sélectionné au moins une fois par l'un des deux programmes dans chacune des trois langues (traduction GT et DL, 25/10/2021) :

	GT	*DL*
Take it to the pens.	Apportez-le aux enclos.	Apportez-le aux stylos.
	Llévalo a los bolígrafos.	Llévalo a los corrales.
	Portalo alle penne.	Portalo alle penne.

Dès 2018, on remarquait cependant que dans la phrase *Take it to the pens near the house*, l'ajout du mot *house*, suffisait pour que le résultat obtenu par les deux logiciels soit correct dans les trois langues :

GT	*DL*
Apportez-le aux enclos près de la maison.	Emmenez-le dans les enclos près de la maison.
Llévelo a los corrales cerca de la casa.	Llévalo a los corrales cerca de la casa.
Portalo nei recinti vicino alla casa.	Portalo nei recinti vicino alla casa.

Certains problèmes subsistent en cas d'ellipse, comme dans le cas d'emploi de *pen* en tant que forme réduite du mot ***play****pen* (dont la traduction correcte serait respectivement *parc* / *corralito* / *box), parc s'entendant ici dans le sens* d'un enclos permettant à un enfant en bas âge de jouer en sécurité et d'apprendre à marcher, comme on le voit dans la

traduction de la phrase *The baby was playing in her pen* (traduction GT et DL, 25/10/2021) :

GT	*DL*
Le bébé jouait dans son enclos.	Le bébé jouait dans son enclos.
La bebé estaba jugando en su pluma.	El bebé estaba jugando en su corral.
Il bambino stava giocando nella sua penna.	La bambina stava giocando nel suo recinto.

Les deux traductions françaises sélectionnent ici le sens correct, mais emploient le terme générique *enclos* au lieu du mot *parc*. Le traitement est similaire en espagnol (ou *corral* et *corralito* sont cependant tous les deux utilisés) et en italien (ou *recinto* ne peut être utilisé dans ces sens), le sens « stylo » étant sélectionné par GT dans les deux langues.

3.2 La polysémie

Les difficultés créées par la polysémie sont également souvent résolues par la TA grâce à l'environnement lexical. Ainsi, dans le cas d'une phrase telle que *Before 2014 some insurance policies would not cover expenses due to pre-existing conditions*, la polysémie du nom *policy* (qui peut signifier *politique* ou *police d'assurance*) est désactivée par la présence du nom prémodificateur *insurance*. En revanche, le mot *conditions* (qui signifie dans ce contexte *maladies* ou *pathologies*) est traduit littéralement (traduction GT et DL, 16/10/2021) :

Before 2014 some insurance policies would not cover expenses due to pre-existing conditions.	*Avant 2014, certaines polices d'assurance ne couvraient pas les dépenses en raison de conditions préexistantes.*	*Avant 2014, certaines polices d'assurance ne couvraient pas les dépenses liées à des conditions préexistantes.*

Remarquons au passage que l'absence d'indentification de la proposition relative réduite (expenses **[that were]** due to pre-existing conditions = les frais qui étaient dus à des pathologies antérieures) est également une source d'erreur dans la traduction proposée par GT. En espagnol, le sens correct (*enfermedades*) est sélectionné par DL, les deux programmes employant en italien *condizioni*, dont l'emploi est possible dans ce sens.

3.3 Concentration de plusieurs termes polysémiques

La concentration de plusieurs polysémes dans la même phrase peut poser des problèmes insurmontables pour la TA en langue de spécialité, notamment quand le contexte intraphrastique est insuffisamment précis. Ainsi, dans une phrase telle que

Changing the placement of beams relative to the staff involves changing the direction of the stems in the beam, un humain peut rapidement déduire que *staff* ne fait pas référence à un groupe de personnes, *beam* et *stem* ne pouvant désigner que des objets. Mais il se trouve que ces trois mots peuvent chacun revêtir de nombreux sens :

Beam : poutre, rayon, faisceau
Stem : tige, queue (feuille, fruit), pied (verre), racine (mot)
Staff : bâton, hampe (drapeau), état-major, personnel

Les difficultés proviennent donc ici de la mutiplicité des sens de ces polysémes monosyllabiques, trois ou quatre pour chacun des mots concernés, auxquels il faut ajouter leur sens spécialisé dans le domaine de l'écriture musicale (respectivement *ligature*, *hampe* et *portée*). Hors de tout contexte, la phrase ne serait d'ailleurs pas forcément compréhensible pour les anglophones qui ne pratiquent pas la musique. En revanche, si l'on précise que le contexte est celui de l'édition musicale, le sens de *staff* devient plus clair pour un humain, et celui de *stem* et de *beam* peut alors être logiquement déduit par un humain, alors que le traitement par la TA de chacune des ambiguïtés sémantiques est nécessairement disjoint, comme on le voit dans la traduction de DL, où seul *beam* est correctement traduit (traduction GT et DL, 17/10/2021) :

*Changing the placement of **beams** relative to the **staff** involves changing the direction of the **stems** in the beam.*	*Changer le placement des faisceaux par rapport au personnel implique de modifier la direction des tiges dans le faisceau.*	*Changer le placement des faisceaux par rapport à la portée implique de changer la direction des tiges dans le faisceau.*

La traduction de *staff* dans le sens *personnel* est présente dans les trois langues pour GT, alors que l'ET correct (*pentagrama*) est également sélectionné par DL en espagnol. En italien, aucun des trois polysèmes n'est traduit correctement :

GT	*DL*
Cambiar la colocación de vigas en relación con el personal implica cambiar la dirección de los tallos en la viga.	El cambio de la colocación de las vigas con respecto al pentagrama implica el cambio de la dirección de los tallos en la viga.
La modifica del posizionamento dei raggi relativi allo staff comporta il cambiamento della direzione degli steli nel raggio.	Cambiare il posizionamento dei fasci rispetto all'asta comporta il cambiamento della direzione degli steli nel fascio.

Disponible en ligne à https://www.asjp.cerist.dz/en/Articles/155

3.4 Homonymie Nom/Verbe due à la conversion

Le changement de catégorie grammaticale sans suffixation (conversion) est l'une des caractéristiques de l'anglais qui favorisent l'ambiguïté syntaxique. Celle-ci peut se manifester de deux manières distinctes dans le cas de la conversion du nom au verbe ou inversement :

- Un verbe à l'infinitif (*cut*) est graphiquement identique au nom au singulier (pour ce verbe irrégulier, s'y ajoutent les formes du *simple past* et du participe passé).
- Un verbe à la troisième personne du présent (*cuts*) est graphiquement identique au nom au pluriel.

À l'époque où la TA utilisait encore les systèmes à base de règles, l'homonymie due à la conversion était une fréquente source de difficultés (Maniez 2008). Voici un exemple d'erreur commise par le programme de TA Systran en 2008 :

*In general, the probability of physical dependence increases as the daily dose and duration of treatment **increase**.*	*Généralement la probabilité de la dépendance physique augmente à mesure que la dose quotidienne et la durée de l'**augmentation** de traitement.*

Dans la traduction de Systran, le verbe *increase* est traduit par un nom (*augmentation*), avec pour résultat une syntaxe incorrecte dans la langue cible. Les traductions de cette phrase sont correctes dans les trois langues et pour les deux logiciels depuis quelques années. Dans les traductions ci-dessous, on observe d'ailleurs une certaine diversité dans la manière dont le verbe *increase* est traduit (traduction GT et DL, 17/10/2021) :

GT	*DL*
En général, la probabilité de dépendance physique augmente avec l'augmentation de la dose quotidienne et de la durée du traitement.	En général, la probabilité d'une dépendance physique augmente avec la dose quotidienne et la durée du traitement.
En general, la probabilidad de dependencia física aumenta a medida que aumenta la dosis diaria y la duración del tratamiento.	En general, la probabilidad de dependencia física aumenta a medida que se incrementa la dosis diaria y la duración del tratamiento.
In generale, la probabilità di dipendenza fisica aumenta all'aumentare della dose giornaliera e della durata del trattamento.	In generale, la probabilità di dipendenza fisica aumenta all'aumentare della dose giornaliera e della durata del trattamento.

La subordonnée *as the daily dose and duration of treatment increases*, exprime un processus progressif (une traduction possible serait « au fur et à mesure qu'augmentent la dose quotidienne et la durée du traitement »). Cette subordonnée est rendue en français par l'utilisation du nom *augmentation* dans un groupe prépositionnel (*avec (l'augmentation de) la dose*), en espagnol par celle d'un verbe conjugué dans une subordonnée (*a medida que aumenta/ se incrementa*) et en italien par l'utilisation du verbe à l'infinitif (*all'aumentare della dose*).

Si les problèmes posés par l'homonymie par conversion semblent désormais bien résolus par la TA neuronale, quelques erreurs peuvent cependant encore se produire dans le cas de phrases longues et syntaxiquement complexes, comme dans l'exemple ci-dessous (traduction GT et DL, 26/10/2021) :

The audit is beset by a paradox: the people who collected these ***remains*** *did so in order to invent "race" as a biological category […] to identify* ***what remains can be considered*** *those of African heritage.*	*L'audit est assailli par un paradoxe : les personnes qui ont collecté ces restes l'ont fait pour inventer la « race » comme une catégorie biologique […] pour identifier* ***ce qui reste peut être considérés*** *comme ceux d'héritage africain.*	*L'audit est confronté à un paradoxe : les personnes qui ont collecté ces restes l'ont fait pour inventer la "race" en tant que catégorie biologique […] pour identifier* ***les restes qui peuvent être considérés*** *comme d'origine africaine.*

La traduction de DL est ici correcte, l'erreur de GT étant due à une double ambiguïté, puisque *what* peut être un pronom ou un déterminant (hors contexte, *what remains* peut donc se traduire par *ce qui reste* ou *quelles dépouilles*). Cette erreur se produit en dépit du fait que le mot *remains* ait été utilisé en tant que nom plus tôt dans la phrase. Les deux traductions espagnoles de la fin de cette phrase sont correctes (*para identificar qué restos pueden ser considerados / considerarse*), mais on observe en italien le même schéma qu'en français, avec une erreur de GT (*ciò che rimane può essere considerati*) et une traduction correcte par DL (*quali resti possono essere considerati*).

3.5 Homonymie Nom/Verbe non due à la conversion

Plus rarement, certains cas d'homonymie entre un nom et un verbe ne sont pas dus à la conversion, mais font intervenir des lexèmes d'étymologie distincte. C'est le cas dans l'exemple ci-dessous:

> Chait's reliance on one research study to tie political sentiments in the Old South to the legacy of slavery founders on the facts of history.

À première vue, beaucoup de lecteurs pensent qu'il manque un verbe conjugué dans la principale de cette phrase, et c'est cette absence apparente qui constitue une source de difficultés pour l'humain. La forme verbale est en fait *founders*, homonyme du nom signifiant *fondateurs* au pluriel (*founder* vient du vieux français *fondrer* qui signifiait « couler », d'abord dans un sens transitif, puis intransitif).

Les erreurs de la TA sont probablement dues au fait que *founder* n'est utilisé comme verbe que dans un cas sur 1000, selon des chiffres obtenus dans le *Contemporary Corpus of American English* (Davies 2009). On remarque cependant une évolution dans la manière dont ce type de difficulté est traité par la TA au cours des dernières années. Ainsi, dans des traductions datant de 2019, l'erreur consistant à traduire le verbe *founder* par un nom était systématique dans les trois langues. Afin de pallier l'absence apparente d'un verbe dans la principale, GT et DL y ajoutaient d'ailleurs le verbe manquant en transformant le nom *reliance* en une traduction du verbe dont il est dérivé (*rely*) :

Chait s'appuyait sur une étude pour lier les sentiments politiques du Vieux Sud à l'héritage des fondateurs de l'esclavage sur les faits de l'histoire.	*Chait s'est appuyé sur une étude de recherche pour établir un lien entre les sentiments politiques dans le Vieux-Sud et l'héritage des fondateurs de l'esclavage sur les faits de l'histoire.*

Dans les traductions espagnoles et italiennes, *reliance* était traduit par un nom, donnant ainsi une structure syntaxique fautive, sauf dans la traduction italienne de GT, dans laquelle un verbe (*si affidare*) est utilisé, comme dans les deux traductions françaises :

*La **confianza** de Chait en un estudio de investigación para vincular los sentimientos políticos en el Viejo Sur con el legado de los fundadores de la esclavitud sobre los hechos de la historia.*	*La **dependencia** de Chait de un estudio de investigación para vincular los sentimientos políticos en el Viejo Sur con el legado de los fundadores de la esclavitud sobre los hechos de la historia.*
*Chait **si affida** a uno studio di ricerca per legare sentimenti politici nel Vecchio Sud all'eredità dei fondatori della schiavitù sui fatti della storia.*	La **fiducia** di Chait in uno studio di ricerca per legare i sentimenti politici del Vecchio Sud all'eredità dei fondatori della schiavitù sui fatti della storia.

Disponible en ligne à https://www.asjp.cerist.dz/en/Articles/155

Plus récemment, les traductions obtenues mettent en évidence un processus de segmentation morphémique, les vecteurs de mots utilisés en TA pouvant représenter uniquement certaines chaînes de caractères de ces mots. C'est visiblement ce qui se produit dans le cas du verbe *founder*, pour la traduction duquel on voit apparaître chez DL une forme verbale à partir de 2021, GT continuant à le traduire par une forme nominale (*fondateurs* / *fundatores* / *fondatori*) :

Le fait que Chait s'appuie sur une étude de recherche pour lier les sentiments politiques dans le vieux Sud à l'héritage de l'esclavage se fonde **sur les faits de l'histoire.**	**La confianza de Chait en un estudio de investigación para relacionar los sentimientos políticos del Viejo Sur con el legado de la esclavitud** fundamenta **los hechos de la historia.**	**L'affidamento di Chait su uno studio di ricerca per legare i sentimenti politici nel vecchio Sud all'eredità della schiavitù** fonda **sui fatti della storia.**

En TA neuronale, en cas d'emploi de mots inconnus ou peu fréquents, c'est désormais une pratique courante que de traduire en se fondant sur les morphèmes dont ces mots sont composés (Sennrich et al., 2015, Salesky et al., 2020). Ici, *founder* semble en effet être interprété comme *found* suivi du suffixe *–er*., d'où l'utilisation de verbes correspondant à ce sens (*se fonder* / *fundamentar* / *fondare*).

4. Conclusion

Les progrès accomplis par la TA neuronale depuis quelques années sont impressionnants et concourent à la généralisation de la pratique de la post-édition pour un nombre croissant de types de textes. Pour la traduction de l'anglais vers les langues romanes étudiées ici, la majorité des structures syntaxiques potentiellement ambiguës sont interprétées correctement par la TA neuronale, les quelques erreurs rencontrées concernant majoritairement l'absence de prise en compte de la prémodification adjectivale dans les dépendances syntaxiques à distance. Les groupes nominaux d'emploi le plus fréquent sont ceux dont la traduction par la TA est la plus fiable, indépendamment de leur longueur et de leur éventuelle complexité syntaxique.

En revanche, la désambiguïsation sémantique, si elle s'effectue correctement dans les phrases longues et quand les polysèmes sont inclus dans des unités lexicales de taille supérieure, reste difficile dans le cas des polysèmes brefs employés dans un sens rare et sans leurs collocatifs les plus fréquents. Par ailleurs, le processus de segmentation morphémique des mots rares ou inconnus reste une source d'erreur pour la TA.

Dans la mesure où la post-édition continue de se généraliser avec la progression de la TA, son enseignement gagne progressivement les formations des futurs traducteurs. Il importe donc de leur enseigner la détection des difficultés rencontrées par la TA face aux ambiguïtés syntaxiques et sémantiques, ainsi qu'aux autres spécificités des erreurs rencontrées dans les résultats proposés par celle-ci.

Références

[1] Brun, C., Jacquemin, B., & Segond, F. (2005). « Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique lexicale. » arXiv preprint *cs/0506049*.

[2] Davies, Mark. 2009. The 385+ million-word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*(2), 159-190.

[3] Kilgarriff, A., & Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities*, *34*(1), 15-48.

[4] Li, X., Zhang, J., & Zong, C. (2016). Towards Zero Unknown Word in Neural Machine Translation. In *IJCAI* (pp. 2852-2858).

[5] Maniez, F. (2008). « Traduction automatique et ambiguïté syntaxique : le cas de la coordination dans les groupes nominaux complexes en anglais médical » in HEIDEN, Serge et Bénédicte PINCEMIN (dir.) JADT 2008 : *Actes des 9es journées internationales d'analyse statistique des données textuelles*, Lyon, 12-14 mars 2008, pp. 765-776. Lyon : Presses universitaires de Lyon.

[6] Maniez F. & Villar-Diaz M.-B. (2018). « Le traitement des expressions nominales potentiellement ambiguës en traduction automatique et ses conséquences sur l'enseignement de la post-édition », *Actes de l'IVe Colloque international franco-espagnol* (Écrire, lire, traduire, enseigner les langues à l'ère du numérique), Université de Grenade.

[7] Michiels, A. (1999). An experiment in translation selection and word sense discrimination. Thinking English Grammar, *Orbis/Supplementa*, 12, 383-407.

[8] Robert, A.-M. (2010). « La post-édition : l'avenir incontournable du traducteur ? », *Traduire*, n° 222, 2010, p. 137-144.

[9] Salesky, E., Runge, A., Coda, A., Niehues, J., & Neubig, G. (2020). Optimizing segmentation granularity for neural machine translation. *Machine Translation*, 34(1), 41-59.

[10] Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

[11] Tiedemann, J., & Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation, DISCOMT'17*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

[12] Voita, E., Sennrich, R., & Titov, I. (2019). When a good translation is wrong in

context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. *arXiv preprint arXiv:1905.05979.*

Notice biographique

François Maniez est professeur émérite de linguistique anglaise et a enseigné l'anglais de spécialité à l'Université Lumière Lyon 2 (France), où il a dirigé le Centre de Recherche en Terminologie et Traduction de 2007 à 2017. Ses principaux domaines de recherche sont la linguistique de corpus, la lexicologie, la traduction automatique et l'anglais de spécialité, notamment la syntaxe et le lexique de la langue médicale.