Traduction automatique et diversité des langues et des textes

Jean-Louis Vaxelaire

Journal of Data Mining and Digital Humanities

Traduction automatique et diversité des langues et des textes

Jean-Louis Vaxelaire,

Université de Namur, Belgique

jean-louis.vaxelaire@unamur.be

Résumé
Les bons résultats présentés dans les travaux de recherche tendent à tous être liés à l'anglais, soit en langue source, soit en langue cible. La traduction automatique doit pouvoir fonctionner avec des couples de langues différents, nous observerons donc si le passage du turc au français est suffisant. Nous verrons que le type de texte est important au niveau des résultats : si les dépêches d'agences sont plutôt correctement traduites, ce n'est pas le cas de la majorité des messages laissés sur les réseaux sociaux. Les chercheurs doivent prendre en compte cette diversité des textes et réfléchir à des matériaux d'entraînement plus spécifiques pour obtenir de meilleures traductions.

keywords
Traduction automatique, turc, genres textuels

INTRODUCTION

Le grand public est encore très critique envers la traduction automatique, il n'est pas rare de trouver sur Internet des pages qui affichent des résultats malheureux et souvent très amusants. Heureusement, la TA a, ces dernières années, plus progressé en qualité qu'en réputation : les résultats sur certains textes sont parfois bons, voire excellents. Nombre de personnes qui se moquent des traductions automatiques des années 2000 seraient étonnées par le rendu de quelques articles journalistiques entre l'anglais et le français sur DeepL.

Les travaux de recherche ne sont sans doute pas dirigés en priorité vers le grand public, les besoins de TA sont en fait essentiels pour les institutions internationales, les gouvernements et les entreprises. Le problème qui se pose à nous aujourd'hui est la mauvaise perception qu'ont les décideurs de la diversité linguistique. J'aurais personnellement tendance à défendre les traducteurs humains ou la pratique de l'intercompréhension lorsque les langues sont proches, mais les traducteurs ne sont pas toujours disponibles et certains jugent que les coûts de traduction sont trop élevés. Seules deux solutions « économiques » se présentent alors à nous : une langue commune, qui est le plus souvent le globish, ou la traduction automatique. En raison des progrès de cette dernière, elle semble une moins mauvaise solution qu'un code restreint qui n'est la langue de personne et qui laisse de côté toutes les nuances d'une langue authentique.

Toutefois, malgré les progrès, il demeure de nombreux points problématiques au niveau de la TA, comme par exemple la question des types de langues, que nous aborderons dans la première partie avec le cas du turc, une langue dont le fonctionnement est très éloigné des langues romanes et germaniques. Puisque les résultats ne sont pas toujours probants, nous nous arrêterons sur la question du genre qui est essentielle dans la traduction humaine et pourtant peu abordée dans les discussions théoriques. Enfin, dans la dernière partie, nous réfléchirons aux problèmes de méthodes qui entravent l'amélioration de la TA.

I DIVERSITÉ DES LANGUES

ISSN 2416-5999, an open-access journal

D'après un audit de PricewaterhouseCoopers datant de 2018, nous n'aurions plus besoin d'ici 2030 de différentes professions, entre autres les radiologues pour analyser les clichés et les traducteurs pour passer d'une langue à l'autre (Gelin & Guilhem, 2019). S'il ne peut se prononcer sur l'avenir des radiologues, le linguiste va difficilement se laisser entraîner par l'enthousiasme actuel d'équipes de recherche qui affirment avoir touché le saint-Graal, c'est-à-dire obtenir de meilleurs résultats que les traductions humaines. L'histoire de la TA nous apprend que ce n'est pas la première fois que l'on annonce avoir réussi à vaincre la barrière des langues (Vaxelaire, 2021). Un autre bémol me semble nécessaire : on ne peut nier que les succès présents se font toujours avec l'anglais dans l'équation et la TA doit pouvoir exister avec d'autres langues. Un dernier point problématique concerne la traduction en tant que pratique : peu de gens savent réellement ce qu'est la traduction – Combe (2009) le démontre au sujet de l'interprétation – et, en lisant les travaux des informaticiens qui travaillent dans le domaine de la TA, on est en droit de douter que tous voient la traduction autrement que comme le passage d'un code à un autre (par exemple la définition qu'en donnait Weaver après-guerre : le russe est de l'anglais codé différemment). Dans un ouvrage bien plus récent, Kelleher écrit « Machine translation is a function that takes a sentence in one language as input and returns the translation of that sentence in another language. » (2019 : 16). Considérer ainsi la traduction ne permet pas de percevoir la complexité de la tâche, le risque est alors grand d'échouer, ce qui a été le cas de très nombreux projets au fil des décennies.

La TA est un outil qui peut se révéler indispensable aujourd'hui. Comme l'indiquent Nunes Vieira *et al.* (2021 : 1519), on a parfois réussi à traiter des patients grâce à la TA car il s'agissait de la seule solution avec des patients et des soignants qui ne parlent aucune langue en commun. Les risques sont évidents, d'un côté comme de l'autre, mais les médecins en sont conscients. Les auteurs notent par exemple que « your child is fitting » était traduit en swahili par l'équivalent de « votre enfant est mort ». Il ressort aussi de ce travail que l'utilisation de Google Traduction aux urgences donnait 2% d'erreurs potentiellement graves en espagnol et 8% en chinois, ce qui veut donc dire qu'il vaut mieux passer de l'anglais à l'espagnol que de l'anglais au chinois.

Le turc a pour nous l'intérêt d'être une langue agglutinante, c'est-à-dire d'un type différent de l'anglais ou du français, et qui n'a pas de genre, un point évidemment problématique pour la traduction. Puisque cette langue n'existe pas encore sur DeepL, j'ai utilisé Google Traduction[1] (désormais GT), Reverso (désormais Rv), Translator.eu qui utilise la technologie Microsoft (désormais T.eu), Webtran (désormais WT) et divers exemples proviennent de Facebook (désormais FB).

Si dans une langue comme le français, les relations seront marquées par des démonstratifs ou des prépositions, dans une langue agglutinante, des suffixes jouent ce rôle. Ainsi, « une maison » se traduit littéralement par *bir ev*, mais le déterminant disparaît dans « ma maison » qui deviendra *evim*, le suffixe possessif de première personne *–im* est alors ajouté au nom. Ce cas est bien traité par les différents systèmes de TA. Le résultat est également bon pour d'autres exemples comme *evimdeyim* traduit par « je suis à la maison ». Dans ces exemples isolés, l'agglutination ne semble pas poser de problème. Toutefois, des travaux comme ceux d'Ataman *et al.* (2017) indiquent pratiquer un pré-traitement morphologique avant de laisser la machine faire son œuvre[2]. D'après Oflazer & Durgar El-Kahlout (2007), il y a environ 150 suffixes en turc, intégrer cette liste pourrait donc améliorer le rendu.

[1] Pour ne pas alourdir le texte, je n'ai pas précisé à quel moment les exemples ont été testés, tout a été revérifié entre décembre 2021 et janvier 2022. Il est possible que les résultats soient différents à d'autres moments.

[2] Cette question du pré-traitement est à soulever pour d'autres paires de langues comme le souligne Delferrière (2021) à propos des expressions figées entre le français et l'espagnol. Ainsi, être « à couteaux tirés » de *L'Assommoir* devient « los cuchillos desenfundados » sur DeepL que le contexte soit la phrase ou le paragraphe, mais « en desacuerdo » sans contexte et ponctuation. De même, « … pleurait à chaudes larmes » devient

ISSN 2416-5999, an open-access journal

L'homonymie est clairement un obstacle à la TA. Ainsi, si j'écris « Je voudrais une eau française », GT me donne la traduction *Fransız suyu istiyorum*, mais si on retraduit en français, on obtient « Je veux du jus français » car *eau* et *jus* sont un seul et même signifiant en turc. Bien qu'il ne parle pas de TA, Golstein met le doigt sur un problème encore plus spécifique : l'homonymie des suffixes en turc. Ainsi, le suffixe personnel verbal de premier type et le suffixe personnel possessif à la première personne du singulier après une consonne ont la même forme, ce qui donne naissance à une ambiguïté grammaticale. *Şoförüm* peut ainsi être traduit par « je suis chauffeur » et « mon chauffeur » (Golstein, 1999 : 200-201), un contexte plus large est nécessaire pour les distinguer. De même, *atı* peut donner « le cheval » en position objet ou « son cheval » car *-ı* est la marque de l'accusatif mais aussi celle du possessif à la 3e personne. L'analyse est également rendue complexe par un autre type d'homonymie, à l'instar du terme *okulları* qui peut être découpé en *okul-ları* (« leur(s) école(s) » car *-ları* peut être le possessif de la 3e et de la 6e personne) ou en *okul-lar-ı* (« ses écoles » avec une marque du pluriel *-lar* et *-ı*, le possessif dont nous avons parlé plus haut), et sans contexte le choix est impossible. Étrangement, malgré le suffixe du pluriel, Rv et T.eu proposent seulement « école » comme traduction, GT ajoute le pluriel pour obtenir « écoles » et seul WT ajoute un possessif avec « leur école ».

Un exemple du même type fonctionne mieux : *babaları* devient « leur père / leurs pères » pour RV, « leur père » (GT) ou encore « leurs pères » (WT), seul T.eu se fixe sur « père ». Par contre, si l'on ajoute d'autres suffixes comme dans *babalarına*, le possessif apparaît enfin : WT propose « à leur père » alors que GT et WT utilisent le pluriel (« à leurs pères ») et Rv ajoute plusieurs possibilités : « à leur père / à leurs pères / leur papa ».

L'absence de genre du turc est problématique pour les traducteurs humains, il faut un contexte plus large pour savoir si un énoncé comme *Kedisini sattı* doit être traduit par « Il » ou « Elle a vendu son chat ». Nos systèmes y voient un masculin, « Il a vendu son chat », sauf WT qui omet le pronom sujet : « Vendu son chat ». Dans un cas comme *O bir şarkıcı*, il n'est pas non plus possible de deviner le sexe du chanteur en question. GT penche pour un homme (« c'est un chanteur ») alors que les autres estiment qu'il s'agit d'une chanteuse. Avec des métiers plus « sérieux » (juge, banquier, architecte), on note presque toujours le masculin. En ce qui concerne *O bir doktor*, Rv et T.eu y voient un médecin, GT est le seul à donner « Elle est médecin ». Il n'est pas possible de savoir s'il s'agit de statistique, de stéréotypes ou de hasard[3].

L'exemple suivant est un commentaire sur Youtube à propos d'une chanson d'une artiste féminine. Les résultats sont tous peu satisfaisants :

> *İzin verelim ışığın tamamen sönmesine derken, oradaki ışık ile yalnızca kendi sevgisini dile getiriyor aslında, çünkü biliyor artık sevilmediğini. sevgi veya aşk, en başından veya zamanla çeşitli nedenlerle karşılıksız olduğunda yıkıcı bir hisse dönüşüyor.*
>
> GT : Quand il dit que la lumière s'éteigne complètement, il exprime seulement son propre amour avec la lumière là, parce qu'il sait qu'il n'est plus aimé. Lorsque l'amour ou l'amour n'est pas partagé dès le début ou pour diverses raisons au fil du temps, il se transforme en un sentiment destructeur.
>
> Rv : Disons simplement que lorsque vous dites que la lumière s'éteint complètement, elle n'exprime son amour avec la lumière là-bas, parce qu'elle sait qu'elle n'est plus

« llorando lágrimas calientes » chez GT alors que la même expression à l'infinitif est traduite par « sollozar ». En fin de compte, « Hors contexte et sans ponctuation, [DeepL] est capable de mieux traduire 11% des expressions figées qui avaient été classées comme non satisfaisantes. » (2021 : 13).

[3] Le problème est a priori identique avec les autres langues qui ne possèdent pas de genre, ainsi dans un message en polonais sur le décès de l'artiste hip-hop MF Doom, FB choisit le féminin, Rv et T.eu le masculin et GT un neutre (« Cela nous inspirera toujours » au lieu de « Il continuera à nous inspirer »).

ISSN 2416-5999, an open-access journal

aimée. l'amour ou l'amour est un sentiment dévastateur lorsqu'il n'est pas partagé depuis le début ou au fil du temps pour diverses raisons.

T.eu : Laissons simplement la lumière s'échapper complètement, mais avec la lumière là-bas, il n'exprime que son propre amour, parce qu'il sait qu'il n'est plus aimé. l'amour ou l'amour devient un sentiment destructeur dès le début ou au fil du temps lorsqu'il reste sans réponse pour diverses raisons.

WT : Laissons la lumière est complètement éteinte par la lumière avec la lumière Il n'y a que le fait qu'il exprime son propre amour, car il n'est plus aimé. L'amour ou l'amour tournent une part destructrice quand il est non assiégé pour diverses raisons du début ou du temps.

Il n'y a que Rv pour donner le bon genre car tous les autres adoptent le masculin, mais rien dans cet extrait ne permet de savoir si l'on parle d'un homme ou d'une femme. On voit à nouveau le problème des termes proches avec *sevgi veya aşk* qui est à chaque fois rendu par « l'amour ou l'amour » alors qu'il est évident que l'auteur ne les considère pas comme des synonymes (*sevgi* est plus proche de la tendresse, on peut employer ce terme avec ses parents, son enfant, alors que *aşk* est un sentiment plus fort). Enfin, divers problèmes grammaticaux rendent difficiles la lecture de certaines de ces traductions, mais il serait trop long de s'y attarder.

Si certaines erreurs sont compréhensibles, d'autres le sont bien moins. Avec *Ali bey de ben de sigara içeriz*, GT et T.eu présentent une bonne traduction : « M. Ali et moi fumons tous les deux. » À l'inverse, Rv donne « Ali bey et moi allons fumer », en ne modifiant pas *bey* (qui certes existe en français, mais pas dans cette acception) et en ajoutant un futur proche qui n'est pas utile dans ce contexte. La traduction de WT est encore plus curieuse car dans « Ali Bey, je fume aussi », le sujet est changé alors que la terminaison *-iz* du verbe ne marque que la 4^{e} personne, le passage à la 1^{e} est énigmatique.

Certaines ambiguïtés ont peu de risques de se produire dans un contexte complet, mais malheureusement celui-ci n'est pas une réalité pour la TA . Ainsi, dans l'exemple *Gel güzelim*, on sait généralement quel est le genre de la personne concernée si l'on a accès à l'intégralité du texte : s'il s'agit d'un « chéri » pour GT, c'est une « chérie » ou une « belle » pour les autres systèmes. Étrangement, la présence ou non d'une virgule peut modifier les résultats. Si Rv garde dans les deux cas « Allez, ma belle, on y va », T.eu change « Allez, chérie » pour *Gel güzelim* en « Allez, bébé » pour *Gel, güzelim* avec virgule et GT modifie le genre de la personne : le « Viens chéri » initial devient « Viens, chérie ». Enfin, WT transforme « Viens ma belle » en « Viens, je suis belle[4] » : personne ne pensait qu'une virgule pouvait avoir de telles conséquences.

Le turc peut employer la répétition pour intensifier le propos. Par exemple, *fısıl fısıl* est correctement traduit par GT (« à voix basse »). WT propose « en murmure » et T.eu bégaye « chuchoter chuchoter » (alors que le verbe « murmurer » est donné lorsqu'il n'y a qu'une occurrence de *fısıl*). Rv est de son côté parti dans une mauvaise direction avec le verbe « brancher[5] ». Un autre exemple de répétition est donné par Guise (2014) : *Çıldır çıldır*, qu'il traduit par « brightly, with a sparkle ». Cette fois-ci, c'est GT qui bégaye : « deviens fou deviens fou ». Quant à WT, on voit apparaître de l'anglais : « Go crazy est fou ».

Le problème de la nuance n'est pas nouveau pour la traduction automatique : nous avons tous vu une liste de verbes proches qui étaient tous traduits par le même terme dans la langue-cible, nous avons aussi vu précédemment qu'aucun de nos traducteurs ne permettait de distinguer les noms *sevgi* et *aşk*. On est en droit de se demander si le passage par une langue-pivot n'accentue pas le problème. Pour le turc, prenons trois cas : *fısıltı hâlinde, fısıldayarak, alçak sesle*. GT ne les distingue pas et propose trois fois « à voix basse ». Pour T.eu, c'est

[4] La confusion relève du même problème d'homonymie que *şoförüm*.

[5] Avec une seule occurrence, on demeure dans le domaine de l'électricité : c'est le nom « fiche » qui apparaît.

ISSN 2416-5999, an open-access journal

« chuchoter, chuchoter, bas », « chuchote, chuchote, crie » pour Rv avec un problème évident pour le dernier. Seul WT donne trois possibilités : « dans le murmure, murmurant, voix basse ».

Sağ (2021) met l'accent sur une particularité grammaticale du turc, le fait que la marque du pluriel puisse être absente dans certains cas. Ainsi, dans l'exemple *Ali kitabı okudu*, le terme *kitap* (« livre ») a la marque de l'accusatif mais pas le suffixe *-lar*. Pourtant, dans cette position d'objet direct (la phrase turque adopte le modèle S-O-V), il peut référer à plusieurs éléments : Ali a lu un ou plusieurs livres selon le contexte. Le même phénomène se retrouve juste avant la copule *var* : *Odada fare var* implique qu'il y a une ou plusieurs souris dans la pièce alors que *fare* se présente comme un singulier. Enfin, Sağ propose un troisième cas, lorsque le nom est en position de prédicat, par exemple dans *Ali ve Merve çocuk*, qui est nettement moins ambigu : littéralement nous avons *Ali + et + Merve + enfant*, il ne fait donc aucun doute qu'il s'agit de traduire par « Ali et Merve sont des enfants. »

Il me semblait que ce dernier cas ne serait pas problématique puisque GT et WT donnent également cette traduction[6], il y a pourtant des problèmes avec T.eu qui ne semble pas percevoir que *Merve* est un prénom (la majuscule disparaît) et donne : « Ali et enfant merve. » Enfin, Rv passe par l'anglais et propose « Ali et Merve Boy. » En ce qui concerne *Ali kitabı okudu*, GT, Rv et T.eu se contentent du singulier « Ali a lu le livre » et WT fait une erreur en donnant un impératif qui n'a rien de logique : « Ali lisez le livre ». Enfin, le singulier est aussi de mise pour *Odada fare var* qui devient « Il y a une souris dans la pièce » (GT et Rv), « « Il y a une souris dans la chambre » (WT) et « Il y a un rat dans la pièce » (T.eu).

Un exemple supplémentaire me paraît utile : dans *Ben genelde ev tasarlatım araba değil*, il est évident pour tous les humains que, dans ce contexte où le locuteur dit qu'il conçoit habituellement des maisons et non des voitures, les noms *ev* et *araba* sont inévitablement perçus comme étant au pluriel. T.eu et GT s'arrêtent au singulier avec respectivement « J'ai l'habitude de concevoir une maison, pas une voiture » et « Je conçois généralement une maison, pas une voiture ». Rv passe au pluriel mais fait un contresens et oublie les voitures : « D'habitude, je ne dessine pas les maisons ». Enfin, WT se perd totalement en mélangeant les pièces du puzzle : « Je ne suis généralement pas une voiture de conception à domicile ». Tous les traducteurs humains voient intuitivement le pluriel, même s'il n'est pas marqué. L'intuition manque du côté des systèmes de TA…

Guise (2014) met en avant une autre particularité du turc qui est l'absence de déterminant défini en position sujet. Là où en anglais, on fera la différence entre emploi général (*tea is expensive*) et emploi particulier (*the tea is cold*), il n'y aura aucun article dans les deux cas en turc (*çay pahalı* et *çay soğuk*). Parce que la comparaison n'aurait pas de sens en français, j'ai utilisé l'anglais comme langue-cible. Cette différence est parfaitement gérée par GT, Rv et T.eu, seul WT ajoute un article dans « The tea is expensive ».

Parmi les résultats satisfaisants, on peut noter le cas de *çok* qui signifie « très », mais qui peut vouloir dire « trop » dans certains contextes. Ainsi *Onu almadım, çok pahalıydı* est bien rendu par « Je ne l'ai pas pris, c'était trop cher » par GT, T.eu et Rv, seul WT reste trop littéral avec « Je ne l'ai pas pris, c'était très cher. »

Nous reviendrons sur la question du genre textuel dans la prochaine partie, mais nous pouvons déjà affirmer que les styles populaires ou vulgaires posent souvent de gros problèmes. Dans son roman *Tırpan*, Fakir Baykurt écrit : *Tabanlarımın altını öpeymiş! Bokumu yiyeymiş! Serseri!*

GT : Il a embrassé le bas de mes semelles ! Il a mangé ma merde ! Coquin!
WT : Embrasser le fond de mes semelles! La merde de la merde! Coquin!
Rv : Il s'est embrassé sous mes armes ! Il mange ma merde ! Punk !

[6] Par contre, si l'on insère une erreur dans l'énoncé-source (*Ali ve Merve bir çocuk.), elle n'est pas corrigée par GT : « Ali et Merve sont un enfant. »

ISSN 2416-5999, an open-access journal

T.eu : Il m'a embrassé sous mes semelles! Il est foutu ! C'est un clochard !

Dans les dictionnaires bilingues, on trouve comme traduction de *serseri* les termes « vagabond, va-nu-pieds, clochard, rodeur, truand ». La présence de « punk » sur Rv fait alors penser à un passage par l'anglais, puisque dans cette langue ce terme signifie aussi « vaurien ». La traduction littérale de la première partie implique que les expressions imagées sont difficiles à traduire automatiquement. « Embrasser le bas des semelles » en turc, c'est en fait supplier quelqu'un pour se faire pardonner, mais aucun système n'arrive à se détacher des mots de la langue source. La traduction de *Bokumu yiyeymiş* par GT est sans doute la plus parlante, mais « coquin » a un côté suranné qui est étrange dans ce contexte.

Les expressions figées souffrent également d'une trop grande littéralité. *Balık baştan kokar* est la version turque de l'adage d'Érasme, « le poisson pourrit par la tête ». Pour GT, « le poisson pue de la tête », T.eu oublie la préposition (« « le poisson pue la tête »), Rv choisit un mauvais homonyme (« le poisson pue depuis le début ») et WT donne comme souvent une traduction inutilisable[7] (« le poisson sent le début »).

Je ne prendrai qu'un seul exemple dans le sens français-turc, celui d'une métonymie. Le français a plus souvent recours à la métonymie que d'autres langues, mais si cette figure est facilement compréhensible en contexte, elle l'est moins dès que l'on traite des syntagmes ou des phrases isolés. Ainsi, ce titre de *Libération* « Les robes noires virent au jaune » (11/01/19) pourrait juste poser une petite difficulté si l'on souhaite conserver une forme de jeu de mot. Puisque l'article apparaît dans la rubrique « justice », il est évident pour les lecteurs que l'on parle d'avocats avec ce terme de *robes noires*. Le *jaune* du titre fait référence aux crises sociales qui touchent alors la France (le mouvement des gilets jaunes). La traduction littérale donnée par GT, Rv et T.eu (*Siyah elbiseler sarıya döner*) n'est pas interprétable par les turcophones et celle de WT (*Siyah elbiseler sarı döndü*, ce qui veut plutôt dire « les robes noires sont devenues jaunes ») l'est encore moins. Si la métonymie n'existe pas en turc[8], la traduction littérale n'amène que des incompréhensions dans la langue cible.

Pellentesque dignissim ultrices fringilla. Vivamus eu luctus ante, vel bibendum magna. Curabitur elit purus, tincidunt non dui vitae, elementum bibendum neque. Curabitur ullamcorper sit amet justo at hendrerit. Fusce ut arcu imperdiet nibh mollis tempus a aliquet tellus. Quisque pharetra cursus nisi, vel lobortis ante consectetur et. Vivamus sed congue neque. Proin pellentesque risus nec dui consequat rutrum. Vestibulum nunc diam, placerat quis auctor vel, faucibus non justo. Etiam dictum purus neque. Phasellus imperdiet mauris ligula, eu laoreet nisi elementum ut. Sed sed porta massa. Aenean faucibus risus ultrices ornare porta. Quisque faucibus ante a tincidunt vestibulum. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

II GENRES TEXTUELS

Si les résultats sont meilleurs avec des textes écrits dans une langue plutôt neutre que pour des exemples comme la citation de Baykurt, cela devrait amener les concepteurs à se poser une question : traduit-on de la même manière tous les textes ?

Ce n'est évidemment pas le cas, le critère du genre textuel n'est pourtant jamais pris en compte. Plusieurs articles parlent de conférence TED comme matériau d'entraînement ; c'est une idée excellente s'il s'agit de traduire des conférences TED, il n'est pas certain que cela fonctionne aussi bien avec d'autres types de textes. Gérard (2019) a bien mis en évidence que nous n'écrivons pas de la même manière selon le genre textuel, les traducteurs humains en

[7] Le problème se retrouve dans la version anglaise de l'expression, « The fish rots from the head », qui est bien traduite par GT et compréhensible chez Rv et T.eu, mais pas chez WT : « les robots de poisson de la tête ».

[8] Kemal Dorum, que je tiens à remercier pour son aide, me signale dans une communication privée que des métonymies peuvent être identiques, par exemple *beyaz önlüklüler*, « les blouses blanches », pour les médecins.

ISSN 2416-5999, an open-access journal

sont pleinement conscients, c'est moins le cas des informaticiens qui gèrent les programmes de TA. Poibeau avait pourtant souligné le problème en écrivant que « les textes sur ces réseaux (tweets, story d'Instagram, messages échangés sur Facebook…) sont écrits dans une langue très différente de celles que l'on trouve sur le Web en général. » (Poibeau, 2018 : 142). Par exemple, on y retrouve des emplois non standards, l'utilisation de la première personne ou des invectives qui sont absents des matériaux d'entraînement (*ibid.* : 147). Il suffit d'aller sur Facebook et d'avoir des amis qui parlent des langues différentes pour en voir la conséquence dans les traductions qui remplacent les messages originaux : le turc qui est utilisé par diverses personnes donne des résultats souvent incompréhensibles en français ou en anglais car les textes ne sont pas toujours bien ponctués, bien construits et sans aucune faute d'orthographe.
Ces textes peuvent aussi mélanger les langues, ce qui entraîne une vraie difficulté pour la TA. Par exemple, une association de promotion de la culture française annonce en turc que la fête de la musique sera annulée à cause de la pandémie et termine en employant le français : « ACFCT vous aime. Bonne fête de la musique » La TA de FB traduit tout en anglais, sans qu'on puisse savoir que deux langues ont été employées, sauf à aller vérifier le message original[9]. Cet arasement linguistique est problématique car s'il y a deux langues dans le texte source, ce n'est pas un hasard, on doit donc en retrouver deux dans le texte cible. Dans l'exemple *Defol git telefonumdan bye bye*, FB propose en français « Éloignez-vous de mon téléphone au revoir » et GT « Enlève mon téléphone au revoir ». À l'inverse, Rv (« Lâche mon téléphone, bye-bye ») et T.eu (« Sortir de mon téléphone bye bye ») conservent les mots en anglais dans la version cible.
Si le genre est un critère essentiel pour la qualité, les traductions de textes qui sont plus ou moins du même genre que le corpus d'entraînement devraient être meilleures. Faisons l'essai avec une dépêche Reuters :

> Fransız otomotiv üreticisi Peugeot'nun CEO'su Jean Phillippe Colin, krize rağmen Türkiye'ye ilişkin yatırım projelerinin devam ettiğini, kompakt sedan sınıfı araçların Türkiye'de üretimi için fizibilite çalışmaları yaptıklarını söyledi.
>
> GT : Jean Phillippe Colin, PDG du constructeur automobile français Peugeot, a déclaré que les projets d'investissement en Turquie se poursuivent malgré la crise et qu'ils mènent des études de faisabilité pour la production de véhicules compacts de classe berline en Turquie.
>
> WT : Fabricant de l'automobile français CEO de Peugeot Jean Phillippe Colin, malgré la crise, malgré la crise, malgré la crise, des projets d'investissement continus liés à la Turquie, des véhicules de classe compacts en Turquie en Turquie, en Turquie, ont déclaré avoir des études de faisabilité pour la production en Turquie.
>
> Rv : Jean Phillippe Colin, PDG du fabricant automobile français Peugeot, a déclaré que les projets d'investissement pour la Turquie se poursuivent malgré la crise, et que les véhicules de classe berline compacte font des études de faisabilité pour la production en Turquie.
>
> T.eu : Jean Phillippe Colin, PDG de Français constructeur automobile Peugeot, a déclaré que malgré la crise, les projets d'investissement en Turquie se poursuivent et qu'ils mènent des études de faisabilité pour la production de véhicules compacts de classe berline en Turquie.

La traduction de WT est catastrophique, mais les autres, malgré quelques défauts (les véhicules qui « font des études de faisabilité » pour Rv ou le « PDG de Français » de T.eu),

[9] Ce problème n'est pas exclusif à la TA. Ainsi, lors de la visite du pape en Irak en mars 2021, son discours avait été en italien et se terminait par : « Salam, salam, salam ». Dans les sous-titres, du journal de France 2, on lisait : « La paix, la paix, la paix », sans indiquer que ces mots avaient été prononcés en arabe, ce qui était pourtant important dans ce contexte.

sont compréhensibles et permettent de saisir la dépêche. Par rapport aux autres essais, celui-ci est bien plus concluant.

Il me semble intéressant de présenter un autre exemple de dépêche qui contient une difficulté, la manière dont sont écrits les chiffres en turc. Par exemple, 392 168 s'écrit *392 bin 168*, *bin* signifiant *mille* :

> *Fransa Halk Sağlığı Kurumunun açıkladığı verilere göre, son 24 saatte 268 kişinin yaşamını yitirmesiyle ülkede virüse bağlı can kaybı sayısı 130 bin 15'e çıktı.*
> *Ülkede bir günde 392 bin 168 kişide virüs tespit edildi. Kovid-19 testlerinin pozitif çıkma oranı yüzde 32,8'e yükseldi.*
> *Hastanelerde 3 bin 694'ü yoğun bakımda olmak üzere 30 bin 982 kişinin tedavisine devam ediliyor. Son 7 günde hastanelere tedavi altına alınanların sayısı yüzde 11,19 arttı.* (Anadolu Ajansı, 28/01/21)
>
> GT : Selon les données annoncées par l'Agence française de santé publique, le nombre de décès dus au virus dans le pays est passé à 130 mille 15, avec 268 décès au cours des dernières 24 heures. Le virus a été détecté chez 392 mille 168 personnes en une journée dans le pays. Le taux de tests positifs au Kovid-19 est passé à 32,8 %. Le traitement de 30 982 personnes se poursuit dans les hôpitaux, dont 3 694 sont en soins intensifs. Au cours des 7 derniers jours, le nombre de personnes soignées dans les hôpitaux a augmenté de 11,19 %.
>
> Rv : Selon les données publiées par l'Agence française de santé publique, le nombre de décès liés au virus a augmenté pour atteindre 130 mille 15 au cours des dernières 24 heures, avec 268 décès.
> En un jour, 392 personnes dans le pays ont détecté des virus chez 168 personnes. Le taux positif des tests de Kovid-19 a augmenté à 32,8 pour cent.
> Le traitement de 982 personnes, dont 3 694 en soins intensifs, 30 se poursuit dans les hôpitaux. Au cours des 7 derniers jours, le nombre de personnes qui ont été traitées à l'hôpital a augmenté de 11,19 pour cent.
>
> T.eu : Selon les données publiées par l'Agence française de la santé publique, 268 personnes sont mortes au cours des dernières 24 heures, ce qui porte à 130 015 le nombre de décès liés au virus dans le pays.
> Le virus a été détecté chez 392 168 personnes en une journée. Le taux de tests positifs au covid-19 est passé à 32,8%.
> Les hôpitaux traitent 30 982 personnes, dont 3 694 en soins intensifs. Au cours des 7 derniers jours, le nombre de personnes traitées dans les hôpitaux a augmenté de 11,19%.
>
> WT : Selon les données annoncées par l'établissement de santé publique de France, le nombre de pertes de vie dans le pays avec la vie de 268 personnes au cours des dernières 24 heures a augmenté à 130 000 15.
> Le pays a été détecté dans 392 mille 168 personnes en une journée. Le taux de sortie positif des tests KOVID-19 a augmenté à 32,8%.
> 3 000 694 dans les hôpitaux continuent de traiter 30 mille 982 personnes dans des soins intensifs. Le nombre de traités aux hôpitaux au cours des 7 derniers jours a augmenté de 11,19%.

La traduction donnée par T.eu est réussie, celle de GT est globalement bonne, les chiffres qui contiennent *bin* sont bien traités à la fin, mais pas au début. À l'inverse, Rv a du mal à gérer ces chiffres, ce qui entraîne des erreurs de découpage syntactico-sémantique, comme dans « 392 personnes dans le pays ont détecté des virus chez 168 personnes ». Enfin, comme avec le cas précédent, WT se perd et offre un texte inutilisable. D'autres essais qui ne seront pas abordés ici mènent à penser que les textes de type journalistique sont globalement mieux traités que des textes littéraires ou ceux présents sur les réseaux sociaux.

III QUESTIONS DE MÉTHODOLOGIE

Dans un article du *Huffington Post* de 2011[10], Kurzweil – qui deviendra ensuite directeur de l'ingénierie de Google – prédisait que la TA arriverait à la même qualité que la traduction humaine en 2029. Le tournant neuronal semble avoir rapproché cette date pour de nombreux chercheurs :

> Bidirectional transformers approach estimated human performance for this task on test sets derived by round trip MT on Wikipedia text. They approach, or surpass human performance for this task. (Lappin, 2021 : 92)

Plusieurs questions me semblent importantes : tout d'abord, de quelles langues parle-t-on ? Poibeau insiste sur le fait que les plongements de mots fonctionnent bien avec une « langue à morphologie pauvre avec beaucoup de données disponibles », mais moins pour « les langues à morphologie riche » (2018 : 145). La présence de l'anglais – langue qui correspond parfaitement au portrait-robot du premier type – dans les bons résultats n'est alors pas surprenante. À l'inverse, les résultats peuvent être mauvais pour les langues agglutinantes ou les langues à morphologie riche, à l'instar des langues slaves (*ibid.* : 165).

La question suivante est : de quels humains parle-t-on ? Hassan *et al.* (2018 : 21) comparent leur système à des amateurs. Il serait pour le moins étrange de faire traduire un article scientifique par un neveu ou un voisin qui est en première année de japonais. La traduction est une affaire sérieuse qu'il vaut mieux laisser à des personnes aguerries. J'imagine que les ordinateurs ont rapidement battu des amateurs aux échecs, mais ils ont mis plusieurs années avant de réussir à vaincre des grands maîtres. Aucun système de TA n'est en mesure actuellement de se comparer avec un traducteur professionnel de renom.

Du point de vue universitaire, un bon traducteur est avant tout une personne qui a été bien formée par ses enseignants, traductologues et/ou traducteurs professionnels. Un parallèle avec l'entrainement de la TA pourrait être établi : il est préférable que le système ait appris avec un corpus de qualité si l'on souhaite avoir un bon résultat. Il est nécessaire de se poser la question de l'augmentation de la taille des corpus. À vouloir obtenir des corpus toujours plus importants, la recherche ne se tire-t-elle pas une balle dans le pied ? On trouve de la TA sur de nombreux sites Web et, en allant chercher des corpus parallèles, on tombe inévitablement sur des TA pas toujours fiables. Ainsi, sur glosbe.com, dans une série d'exemples entre le turc et le français, on recense un cas qui provient du site de sous-titres Opensubtitles : *Ve ona sormadan* ***suyu*** *da bozmamalıydın* est traduit par « Et vous avez eu le capitaine d**e** ne pas creuser », ce qui n'a aucun rapport avec l'original. L'idée est particulièrement mauvaise, quand bien même la version française aurait eu du sens car, dans les sous-titres, les traductions peuvent être plutôt libres. Le problème se retrouve aussi sur Linguee, dont la base de données est exploitée par DeepL, où par exemple *a cut above the rest* devient « une coupe au-dessus du repos ».

Les termes techniques ou scientifiques sont un problème pour les traducteurs humains car ils ne sont pas toujours connus et n'apparaissent pas dans tous les dictionnaires. Je m'arrêterai ici sur un nom de papillon en allemand : *Zitronenfalter*. L'original *Wer glaubt, dass ein Abteilungsleiter eine Abteilung leitet, glaubt auch, dass ein Zitronenfalter Zitronen faltet* devient sur FB « Anyone who believes that a department manager leads a department also believes that a lemon butterfly folds lemons. » et n'est même plus un papillon sur DeepL : « Anyone who believes that a head of department runs a department also believes that a lemon folder folds lemons. (DeepL) ». La version de FB semble meilleure, mais si le *lemon butterfly* existe, il s'agit du *Papilio demoleus* alors que le *Zitronenfalter* (qui s'appelle *citron* en français) est le *Gonepteryx rhamni* : le *Papilio demoleus* va souvent sur les citronniers

[10] https://www.huffpost.com/entry/ray-kurzweil-on-translati_b_875745

alors que le *Gonepteryx rhamni* ressemble à une feuille de citronnier. Pour le dire autrement, si l'on veut traduire un texte qui parle de papillons, il est préférable d'avoir un corpus d'entraînement qui traite de papillons. Un corpus trop léger ne peut que poser des problèmes, mais un qui est ciblé sera toujours plus efficace qu'un trop large qui entraîne une forme de bruit.

À quelques exceptions près liées au domaine littéraire, on attend d'un traducteur humain qu'il maîtrise parfaitement les deux langues sur lesquelles il travaille. Il ne viendrait à l'idée de personne de traduire de l'espagnol vers le français en passant par l'anglais. Toutefois, il ne faut pas oublier que pour divers chercheurs, toutes les langues sont équivalentes (de l'anglais codé différemment comme le disait Weaver), l'emploi d'une langue pivot n'est alors pas problématique. Pour prendre une illustration dans un domaine distinct, je peux passer de *quinze* à *XV* en passant par *15* sans que cela ne soit gênant. Je peux d'ailleurs faire l'opération inverse et j'aboutirai toujours à *quinze*. Certes, mais la traduction n'est pas un simple changement de support et n'est pas non plus une retranscription. Il ne s'agit pas non plus de déverbalisation au sens où l'entend la théorie interprétative, mais d'une reverbalisation supplémentaire qui peut entraîner des biais et des erreurs. Nous avons vu plusieurs cas de passage par une langue-pivot entre le turc et le français, on peut trouver la même chose avec l'espagnol sur DeepL : dans un extrait d'un texte littéraire avec une ponctuation peu académique de Goytisolo, *cuando la feraz inventiva del abolido mayo sembraba las paredes de inscripciones burlonas y cínicas, cómicas, feroces, insolentes, dubitativas, alegres* (*Juan sin tierra*) devient « quand l'inventivité fertile des May abolies sème les murs des inscriptions moqueuses et cyniques, comiques, féroces, insolentes, douteuses et cheerful ». Si le manque de corpus entre le turc et le français peut excuser le passage par l'anglais, son utilisation entre l'espagnol et le français est moins soutenable. On peut évidemment se demander si les cas où les deux langues employées dans le texte source sont fusionnées dans le texte cible ne proviennent pas de ce passage par l'anglais.

Un des progrès de la TA neuronale est sans doute de permettre au système de faire des allers et retours dans le texte. Un traducteur humain lit toujours le texte dans son intégralité et sait généralement d'où il provient, il en a donc une vision globale alors que la TA restait jusqu'ici strictement locale. Le contexte est évidemment essentiel pour bien traduire, cela évite les contresens d'un énoncé à l'autre, même si Zheng *et al.* y voient une source de bruit : « The deep hybrid makes the model more sensitive to noise in the context, especially when the context is enlarged. This could explain why previous studies show that enlarging context leads to performance degradation. » (2020). Quand on utilise le terme de *contexte*, c'est aussi pour parler de contexte extralinguistique, ce qui est hors de portée de la TA. Par exemple, *Stretta attorno a Marylin Manson, accusato da più donne di molestie. La rockstar si difende in un breve messaggio* est rendu par FB par « Tighten up around Marylin Manson, accused by multiple women of harassment. Rockstar defends herself in a short message ». Je ne sais pas si *herself* est dû au féminin de *rockstar* en italien ou au prénom *Marilyn*, mais les traducteurs humains auraient utilisé un masculin puisque Marylin Manson est un homme. Le cas d'une pancarte affichée à Florence est plus complexe. Elle est apparue après les rumeurs de départ du meilleur buteur du club local, Vlahovic, chez l'« ennemi » qu'est la Juventus de Turin : *Il rispetto non si conquista con i goal Vlahovic gobbo di merda*. Le texte n'a pas de ponctuation, ce qui aurait aidé DeepL qui peine : « Le respect ne se gagne pas avec des objectifs de bossu Vlahovic » (25/01/22). Le *gobbo* dans ce contexte n'est pas un bossu, mais un supporter ou un joueur de la Juventus. Il est surprenant de voir que « di merda » disparaît dans la traduction, une meilleure version française aurait été « Juventino de merde ». Enfin, *goal* traduit par *objectif* est gênant : l'emploi de l'emprunt à l'anglais montre à n'importe quel être humain qu'il s'agit d'un terme spécifique, en l'occurrence puisqu'il s'agit de football, du mot *but*.

IV POUR CONCLURE

Du point de vue linguistique, il est tentant de résumer succinctement les différentes approches de la TA. Celle qui a dominé la fin du 20e siècle, l'approche par règles, correspondait à une conception affaiblie de la grammaire classique[11], c'est-à-dire à un dictionnaire auquel on ajoutait quelques règles. On est ensuite passé à l'approche statistique grâce au développement de l'informatique. On y reprenait l'idée principale du distributionnalisme qui était de se fonder sur des corpus, avec le défaut identique du rejet du sémantique. L'approche neuronale reprend le principe des corpus, mais on voit pourtant réapparaître des règles, du moins en ce qui concerne les travaux sur le turc. Il est difficile de voir où en est la question sémantique, mais si le système est censé apprendre, cela implique une dimension liée au sens. C'est à ce prix que l'on atteindra de meilleurs résultats.

Pour obtenir une TA plus satisfaisante, il est nécessaire d'avoir une meilleure connaissance de chaque langue (en particulier de ses règles, qui sont différentes de celles des autres) et de prendre en compte les genres textuels. Le traducteur humain est toujours sensible aux questions contextuelles, on ne traduit pas les mêmes suites de lexies de la même manière selon l'auteur ou le genre du texte. Selon Rastier (2021 : 213), il y a trois niveaux de contexte :

- le contexte local (du syntagme à la période) qui permet de déterminer la signification,
- le contexte « textuel » (le contexte, c'est aussi tout le texte), et qui détermine son sens,
- le contexte que constitue le corpus, et qui détermine sa significativité.

Même le premier niveau de contexte est difficilement atteignable pour la TA neuronale, et il s'agit d'accéder au moins au deuxième pour réussir la traduction. La traduction humaine va du global au local car le global détermine le local. Notre problème est que la TA se limite au local, au syntagme ou à la phrase selon les systèmes. Enfin, en ce qui concerne la question sémantique, l'informatique traite des données discrètes alors qu'une langue est difficilement discrétisable, les éléments sémantiques sont diffus dans un texte.

La puissance de calcul et des corpus riches permettent déjà d'avoir de très bons résultats sur certains types de textes, mais les exemples donnés ici montrent qu'il y a encore du travail à accomplir. On peut, comme Kurzweil dans l'interview précédemment citée, dire que même les meilleurs traducteurs humains n'arrivent pas à traduire correctement la littérature[12] pour excuser les lacunes de la TA, mais nous avons observé que le problème existe pour les messages écrits sur Facebook, qui relèvent rarement de la littérature. Il ne s'agit pas de demander à la TA d'atteindre la perfection mais, au minimum, d'éviter les contresens et les passages incompréhensibles. J'ai pris dans un autre article (Vaxelaire, 2021) l'exemple d'une jeune fille d'origine turque qui avait disparu à Vienne : les traductions de l'appel à l'aide de sa sœur ne pouvaient atteindre leur but car elles passaient au masculin et, pour certaines, modifiaient les informations sur le trajet de l'adolescente.

La TA doit prendre en compte le fait que les langues ne sont pas toutes interchangeables et que les textes relèvent de genre qui demandent des stratégies de traduction différentes pour parvenir à de meilleurs résultats.

References

Ataman D. *et al.*. Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*. 2017;108:331-342.

Combe X., *L'anglais de l'Hexagone : constats et réflexions d'un interprète de conférence*, L'Harmattan (Paris), 2009.

Gelin R. & O. Guilhem, *L'IA et nous*, Le Pommier (Paris), 2019.

[11] Qui, de son côté, traitait les langues comme des langues mortes.

[12] Cette affirmation est d'ailleurs fausse, de nombreuses œuvres ont été très bien traduites.

Gérard C. Linguistique des genres : objet et méthode. Statut culturel des genres et variétés génériques. *Lynx*. 2019;18. http://journals.openedition.org/linx/3030.
Golstein B., *Grammaire du turc : Ouvrage pratique à l'usage des francophones*, L'Harmattan (Paris), 1999.
Guise J., *The Turkish Language Explained for English Speakers: A Treatise on the Turkish Language and its Grammar*, auto-édition, 2014.
Hassan H. *et al*. Achieving Human Parity on Automatic Chinese to English News Translation. https://arxiv.org/abs/1803.05567.
Kelleher J.D. *Deep learning*. MIT Press (Cambridge), 2019.
Lappin S. *Deep learning and linguistic representation*. CRC Press (Boca Raton), 2021.
Nunes Vieira L. *et al.*. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*. 2021;24(11):1515-1532. https://doi.org/10.1080/1369118X.2020.1776370.
Oflazer K. & Durgar El-Kahlout I. Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation*. 2007. https://research.sabanciuniv.edu/6397/1/WMT04.pdf.
Poibeau T. *Babel 2.0 : Où va la traduction automatique ?*. Odile Jacob (Paris), 2019.
Sağ, Y. Bare singulars and singularity in Turkish. *Linguistics and Philosophy*. 2021. https://doi.org/10.1007/s10988-021-09323-0.
Vaxelaire J.L. Les progrès de la traduction automatique par le prisme du turc et du luxembourgeois. in J.C. Beacco *et al.* (coord.) : *Traduction automatique et usages sociaux des langues. Quelles conséquences pour la diversité linguistique ?*,. 2021:19-30.
Zheng Z. *et al.*. Toward Making the Most of Context in Neural Machine Translation. 2020. https://www.arxiv-vanity.com/papers/2002.07982/.