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

Traduction d'unités polylexicales du portugais en français par MT@EC et eTranslation

Translation of Multi-Word Units from Portuguese into French by MT@EC and eTranslation

Françoise Bacquelaine
Université de Porto - Portugal
franba@letras.up.pt
Centro de Linguística da Universidade do Porto
0000-0001-8055-1678

Pour citer cet article :
Bacquelaine, F. (2022). Traduction d'unités polylexicales du portugais en français par MT@EC et eTranslation. *Revue Traduction et Langues21* (1), 56-76.

Reçu: 23/05/2022; **Accepté**: 29/08/2022, **Publié:** 31/08/2022

L'auteur correspondant: Françoise Bacquelaine

Keywords

Ambiguity; Corpora; Domain coverage; Neural machine translation; Phraseology; Statistical machine translation

Abstract

*This paper aims to determine the extent to which the shift from statistical machine translation (SMT) to neural machine translation (NMT) improved the performance of European Union machine translation systems between 2015 and 2021 in terms of multi-word unit translation and domain coverage. To do so, we chose to test these systems on machine translation into French of multi-word units expressing quantitative and qualitative progression in Portuguese from Portugal. These units consist of the 2-gram 'cada vez' and a comparative adjective or adverb (cada vez COMP), and their word-for-word translation into French is not idiomatic (*chaque fois COMP). The most frequent translation into French is 'de COMP en COMP'. This implies that these multi-word units must be translated 'en bloc', but their identification is not straightforward. On the one hand, COMP is not fixed and may include one (mais / plus, menos / moins, maior / plus grand, menor / plus petit, melhor / meilleur – mieux, pior / pire – plus mal) or several words (mais or menos N, ADJ, ADV). On the other hand, the 2-gram 'cada vez' can be part of other multi-word units expressing iteration (de cada vez (que)/(à) chaque fois (que)), or 'dropper' ([a certain quantity] de cada vez/ à la fois), This raises the challenge of ambiguity, well known to biotranslators and still often problematic for NMT. Moreover, units expressing quantitative or qualitative progression may raise other translation challenges when they are coordinate (with or without repetition of the 2-gram 'cada vez'), when they are split (cada vez (...) COMP), or when they combine with verbs or nouns to form extended translation units whose translation into French can result in a more concise solution we refer to as 'lexicalisation'. We established a biotranslation model based on a manually aligned French-Portuguese parallel literary corpus and online searchable French-Portuguese aligned corpora (translation memories). We selected a sample of occurrences of these multi-word units including several translation challenges. These occurrences were selected from a Portuguese journalistic corpus. They belong therefore to general language, whereas the EU's translation memories cover the domains dealt with by its institutions, which represents an additional challenge, considering the critical importance of domain coverage in the data to NMT performance quality. The selected occurrences were translated into French by the EU SMT system in 2015 (MT@EC) and 2019 (eTranslation Legacy) and by eTranslation (the EU NMT system) in 2019 and 2021. Firstly, MT output was analysed according to two general criteria: 'non-literality', that is translation into French without 'chaque', and acceptability from a semantic point of view, that is MT output without any false meaning, opposite meaning or nonsense. Then we looked at specific challenges, some of which could lead to original solutions, worthy of a professional human translator, such as lexicalisation, change of grammatical category or 'recategorisation' and 'naturalisation', that is phraseological or syntactic rearrangement that makes the target text more idiomatic. The results show that MT is improving, especially according to the criterion of non-literality. Original solutions are still rare, but they are diversifying in NMT output. Nevertheless, NMT remains imperfect, not least because of the inherent ambiguity of natural languages and the inevitable gaps in the data on which these systems are based. The results also demonstrate the importance of human intervention in the maintenance of the systems learning*

automatically, since the quality of SMT system's output decreases between 2015 and 2019, when all efforts were focused on improving EU NMT system. Finally, results reveal the dangers of using English as a pivot language when translating from one Romance language into another, and the need to train future translators in NMT and post-editing.

Mots clés

Ambiguïté ; Corpus ; Couverture de domaine ; Phraséologie ; Traduction automatique neuronale ; Traduction automatique statistique

Résumé

Cette étude vise à déterminer dans quelle mesure le passage de la traduction automatique statistique (TAS) à la traduction automatique neuronale (TAN) a amélioré les performances des systèmes de traduction automatique de l'Union européenne entre 2015 et 2021 en termes de traduction d'unités polylexicales et de couverture de domaines. Pour ce faire, nous avons choisi de tester ces systèmes sur la TA en français d'unités polylexicales exprimant la progression quantitative et qualitative en portugais du Portugal. Nous avons établi un modèle de biotraduction à partir de corpus parallèles et alignés français-portugais et nous avons sélectionné un échantillon d'occurrences de ces unités polylexicales comportant plusieurs défis de traduction. Ces occurrences ont été prélevées sur un corpus journalistique portugais et soumises aux systèmes de TAS et de TAN de l'UE en 2015, en 2019 et en 2021. Les résultats ont été analysés en fonction de deux critères généraux (non-littéralité et acceptabilité) et de défis particuliers pouvant donner lieu à des solutions originales, dignes d'un biotraducteur professionnel. Il en découle que la TA s'améliore, mais reste imparfaite, notamment en raison de l'ambiguïté inhérente aux langues naturelles et du caractère inéluctablement lacunaire des données sur lesquelles se fondent ces systèmes. Les résultats démontrent aussi l'importance de l'intervention humaine dans l'entretien de ces systèmes, les dangers de l'utilisation de l'anglais comme langue pivot lorsqu'il s'agit de traduire d'une langue romane à une autre et la nécessité d'initier les futurs traducteurs à la TAN et à la post-édition.

1. Introduction

MT@EC et eTranslation sont deux systèmes de traduction automatique (TA) au service des institutions européennes. MT@EC (2013) est un système de TA statistique (TAS) et eTranslation (2017), un système de TA neuronale (TAN). L'objectif de cette recherche est de déterminer dans quelle mesure les efforts réalisés entre 2015 et 2021 pour améliorer ces outils ont porté leurs fruits lorsqu'il s'agit de traduire en bloc des unités polylexicales de la langue générale. En effet, la couverture des domaines est toujours un défi à relever pour la TAN (Koehn, 2020, pp. 294-295 ; Foti, 2021). Or, la spécificité des domaines couverts par les corpus volumineux des services de traduction de l'UE rassemblés dans la base de données Euramis peut constituer un obstacle à la traduction de la langue générale. La Direction générale de la Traduction en est consciente et tente désormais d'y remédier en enrichissant ces données (Foti, 2021).

Nous avons donc soumis un échantillon d'une centaine d'occurrences d'unités

polylexicales exprimant la progression quantitative ou qualitative en portugais à MT@EC en 2015 et à eTranslation en 2019 et en 2021 pour obtenir un corpus de TA brute en français. Ces occurrences en contexte ont été prélevées sur un corpus journalistique portugais de la fin du XXe siècle (CETEMPúblico). L'identification de ces unités polylexicales à traduire en bloc se heurte à plusieurs obstacles : elles peuvent être scindées et elles comportent un comparatif (COMP) variable en plus du 2-gramme *cada vez,* qui entre dans la composition d'autres unités polylexicales, ce qui soulève le défi de l'ambiguïté, bien connu des biotraducteurs et encore souvent problématique pour la TAN (Koehn, 2020, pp. 5-8).

2. Présentation de l'unité de construction préformée *cada vez* COMP

Il est généralement admis que « l'unité de traduction va au-delà du mot, au-delà des mots contigus » et qu'elle résulte de l'unité de sens construite par le traducteur (Durieux, 2014, p. 386). Ainsi, « sur le plan épistémologique, l'unité de traduction n'est pas une donnée stable, telle une unité de sens, elle s'inscrit dans un réseau en constante construction et reconstruction. » (Durieux, 2014, p. 387). C'est le cas des unités polylexicales que nous avons choisi d'analyser ici et qui relèvent, à notre avis de la phraséologie au sens large.

En phraséologie, le concept d'unité de construction préformée (UCP) défini par Schmale (2013) convient bien à l'unité de sens exprimant la progression quantitative ou qualitative en portugais (*cada vez* COMP). En effet, elle répond à la plupart des critères définitoires des UCP de Schmale : elle est polyfactorielle et présente un certain degré de « stabilité ou de figement structural/lexical » permettant « à un membre de la communauté langagière concernée de la reconnaître et de la réutiliser », car elle est fréquente et durable, même si sa longueur est variable et qu'elle n'est pas lexicalement saturée (Durieux, 2014, pp. 41-42).

Certes, son sens est relativement compositionnel, mais sa traduction en français n'est pas littérale. Elle constitue donc une unité de sens à traduire en bloc. Les exemples (1) à (3) proviennent de corpus bilingues de biotraduction[1] et illustrent l'équivalence la plus fréquente entre le français (FR) et le portugais (PT) :

(1) *… a delirante administração comunitária e os danos que ela provoca só podem* […] *dar origem a* ***cada vez mais recursos***.
… la délirante administration communautaire et les méfaits qu'elle engendre ne peuvent […] *qu'engendrer* ***des recours de plus en plus nombreux***. (EP ; LS-FR)

(2) *Sabia* ***cada vez menos*** *o que devia ou não dizer.*
Je savais ***de moins en moins*** *ce que je devais dire ou ne pas dire.* (AN)

[1] EP : Europarl V7 ; LS : langue source ; AN : Amélie Nothomb. Ces corpus sont décrits en 3.1. S'agissant de corpus parallèles de biotraduction professionnelle, les segments alignés sont considérés comme équivalents et nous présentons d'abord la version PT dans tous les exemples.

(3) *Virou as páginas com* ***um frenesim cada vez maior****.*
Il tourna les pages avec ***de plus en plus de frénésie****.* (AN)

L'UCP à traduire en bloc peut se limiter à un 3-gramme, comme en (2), ou être plus étendue, comme en (1) et (3), où *mais* modifie un N. Dans les exemples (4) et (5), le biotraducteur a choisi d'étendre l'unité à traduire en bloc en incluant le V, *há* en (4) et *está a ficar* en (5) pour lexicaliser la progression en FR, en utilisant des formes verbales (*se multiplient, ne cesse de grandir et de se diversifier*) plutôt qu'un COMP, comme dans les exemples précédents, ce qui rend le style plus fluide :

(4) ***Cada vez há mais*** *períodos de extrema seca, seguidos de períodos de grande pluviosidade, …*
Les périodes d'extrême sécheresse suivies de précipitations abondantes ***se multiplient****, …* (EP ; LS-PT)

(5) *A União* ***está a ficar cada vez mais alargada, e cada vez com maior diversidade.***
L'Union ***ne cesse de grandir et de se diversifier****.* (EP ; LS-NL)

L'exemple (4) et la deuxième occurrence de l'exemple (5) illustrent un deuxième défi : l'UCP peut être soudée ou scindée. L'« élément ordonnant » (Leal, 2012, p. 152) ou « agent comptable » (Theissen, 2007, p. 245) *cada vez* peut en effet être séparé de COMP par un – *há* en (4) et *com* en (5) ou plusieurs mots et modifier un ou plusieurs COMP comme dans l'exemple (6) :

(6) ***Cada vez*** *a sociedade da informação depende* ***menos*** *de recursos materiais e de energia,* ***e mais*** *do saber-fazer humano.*
La société de l'information fait ***de moins en moins*** *appel aux ressources matérielles et énergétiques, mais* ***de plus en plus*** *au savoir-faire de l'homme.* (EP ; LS-DE)

L'ambiguïté constitue un autre défi que la machine peine encore à relever (Koehn, 2020, pp. 5-8). Bien qu'il ne s'utilise jamais seul, le 2-gramme *cada vez* est extrêmement fréquent en portugais européen et il est ambigu, car il entre dans la composition d'autres UCP donnant lieu à différentes traductions en français selon que ces UCP expriment la progression, l'itération – *de cada vez* (*chaque fois* ou *à chaque fois*), (*de*) *cada vez que* (*chaque fois que* ou *à chaque fois que*) – ou le 'compte-goutte' – *um de cada vez* (*un à la fois*). Or, le 3-gramme[2] *de cada vez* peut être suivi de COMP :

[2] En traitement automatique des langues, le terme *n-gramme* désigne toute séquence de n éléments (mots, chiffres, signes de ponctuation, …), récurrente dans un corpus, indépendamment de tout autre critère linguistique (voir, par exemple, Granger et Paquot, 2008).

(7) ***Dispomos de cada vez mais conhecimentos*** *sobre as causas profundas de conflitos armados; ...*
Nous comprenons de mieux en mieux *les causes profondes des conflits violents ; ...* (EP ; LS : NL)

On le voit, le traducteur professionnel identifie l'UCP à traduire en bloc sans tomber dans le piège de l'ambiguïté ou de la scission et l'inclut même dans des unités de sens plus étendues qui se superposent à l'unité de traduction (Durieux, 2014, p. 386) pour proposer des lexicalisations de la progression sans COMP.

3. Méthodologie

Pour mener cette étude, nous avons établi un modèle de biotraduction à partir de corpus parallèles et alignés FR-PT et nous avons prélevé un échantillon de test dans un corpus journalistique portugais que nous avons soumis à MT@EC et à eTranslation. La TA brute a été analysée d'après le modèle de biotraduction. L'analyse porte également sur quelques défis particuliers de l'échantillon de test pouvant donner lieu à des solutions originales.

3.1 Modèle de biotraduction

Le modèle de biotraduction de l'UCP *cada vez* COMP a été établi à partir de trois corpus alignés et d'un corpus parallèle. Ces corpus se limitent aux variantes européennes du FR et du PT. L'interface de recherche multilingue du projet OPUS (Tiedermann, 2012) a permis de recueillir les occurrences de l'UCP dans le corpus du Parlement européen (Europarl v7) et de l'Agence européenne des médicaments (EMEA). Le troisième corpus aligné se compose de six sous-corpus sélectionnés en fonction de la variante de PT : un corpus journalistique (PressEurop), l'Acquis de l'UE, Europarl et trois œuvres littéraires[3]. Ces sous-corpus proviennent du corpus tchèque InterCorp (Nádvorníková, 2016) et ont été explorés grâce à l'interface Park. Le roman *Stupeur et tremblements* (Nothomb, 1999) et sa traduction en portugais (*Temor e Tremor*, Almeida, 2000) ont fourni des occurrences alignées manuellement.

Il résulte de l'analyse de ces corpus que l'équivalent FR le plus fréquent de *cada vez* COMP est *de* COMP *en* COMP (ex. 1-3, 6 et 7). Deux autres solutions sont attestées beaucoup plus rarement : *toujours* COMP (8) et COMP (9) :

(8) *... uma vez que a União Europeia tem de se ocupar* ***cada vez*** *de* ***mais*** *matérias, ...*
... l'Union européenne étant censée traiter ***un nombre toujours plus important de*** *domaines, ...* (EP ; LS-PL)

[3] Alice in Wonderland (Lewis Carroll), Harry Potter and the Philosopher's Stone (J. K. Rowling) et The Fellowship of the Ring (J. R. R. Tolkien).

(9) *Há vários estudos que demonstram que as pessoas* ***cada vez*** *vêem* ***menos*** *anúncios, uma vez que podem mudar de canal, …*
Plusieurs études ont montré que les spectateurs regardent ***moins*** *les publicités parce qu'ils changent de chaîne …* (EP ; LS-PT)

Enfin, la lexicalisation de la progression quantitative (4) et qualitative (5) représente une dernière solution de traduction de l'UCP portugaise.

3.2 Échantillon de test

L'échantillon de test provient du CETEMPúblico, un corpus d'environ 180 millions de mots du journal portugais Público couvrant la période de 1991 à 1998. Il s'agit donc de langue générale et de la variété portugaise. Les segments prélevés pour tester les systèmes de TA de l'UE comportent 101 COMP pour 98 occurrences de *cada* vez. Afin de défier la machine, les six COMP portugais y sont attestés, alors que *mais* est de loin le plus fréquent (Bacquelaine, 2020, p. 221), et les éléments modifiés (N, V, ADJ et ADV) sont variés. Nous avons également sélectionné des phrases comportant une succession d'UCP coordonnées, comme en (5), mais aussi sans répétition de *cada vez*, comme en (6), plusieurs UCP élargies (*estar a ficar* ADJ, *tornar-se* ADJ, *estar com* N, *trabalhar* ADV, *acontecer* ADV) et plusieurs constructions du V *saber*, dont la traduction représente un défi et peut donner lieu à une lexicalisation de la progression en FR. La scission est également représentée dans 29 occurrences et une phrase teintée d'humour comporte deux UCP.

Cet échantillon n'est donc pas représentatif de l'usage courant et vise à tester les systèmes de TA sur des cas moins fréquents pour établir jusqu'à quel point la tendance à la généralisation débouche sur des contresens ou des faux sens. Même s'il date du siècle dernier, l'échantillon est pertinent, car l'UCP exprimant la progression n'a pas évolué depuis.

3.3 MT@EC et eTranslation

Le système de TAS MT@EC est fondé sur les corpus alignés de qualité que sont les mémoires de traduction de l'UE (Euramis), expurgés des segments pouvant soulever un problème de protection des données personnelles (Foti, 2021). Il a été développé grâce à l'outil open-source Moses (Koehn *et al.*, 2007) à partir de 2009 pour remplacer le système à base de règles utilisé depuis 1975 (ECMT) et a été mis en place en 2013 (Foti, 2021) au service des institutions européennes et des administrations publiques des États membres, de la Norvège et de l'Islande (Eisele, 2017, p. 6).

Le système de TAN eTranslation a été développé à partir de MT@EC, qu'il a progressivement remplacé à partir de 2017 (Eisele, 2017, pp. 17-19). C'est l'outil open-source Marian NMT (Junczys-Dowmunt *et al.*, 2018) qui a été sélectionné et l'utilisation de ce système de TAN a été élargie aux universités et aux PME européennes (Foti, 2021). La qualité de la TAN dépend des domaines couverts et de la qualité des corpus alignés,

mais ce n'est pas facile de trouver des corpus de qualité en dehors d'Euramis pour couvrir la langue générale et des domaines étrangers à l'UE (Foti, 2021). Des corpus alignés du projet Opus (Tiedermann, 2012) ont été ajoutés aux données. Ainsi, 90 % des données d'eTranslation proviennent d'Euramis et 10 % d'autres corpus alignés (Foti, 2021).

MT@EC et eTranslation nous ont fourni deux échantillons de TAS et deux échantillons de TAN brute. MT@EC a été testé en 2015. En 2019, la TAS était encore disponible sous le nom de *eTranslation Legacy*. eTranslation permet de sélectionner un domaine particulier (santé, finances, droit, …), mais nous avons choisi l'option recommandée pour la langue générale, *Cutting edge* en 2019 et *General Text* en 2021.

3.4 Critères d'analyse de la TA brute

Pour analyser la TA brute, deux critères généraux ont été définis : la non-littéralité, soit la traduction sans *chaque*, et l'acceptabilité. La TA est considérée comme acceptable s'il s'agit d'une des solutions du modèle de biotraduction. Elle est considérée comme inacceptable si la TA constitue un contresens, un faux sens ou un non-sens.

L'analyse porte en outre sur la gestion de la scission et les solutions originales, dignes d'un traducteur professionnel. Nous avons retenu trois types de solutions originales : lexicalisation, recatégorisation et 'naturalisation'. La lexicalisation consiste à remplacer l'UCP élargie par un V, comme dans les exemples (5) et (6) relatifs à la biotraduction, ou par un ADJ dans notre échantillon de TA brute. Le changement de catégorie grammaticale ou 'recatégorisation' (Ballard, 2006) est un procédé classique, qui se passe de commentaire. La naturalisation relève d'un aménagement phraséologique ou syntaxique qui rend le texte cible plus idiomatique et donc plus fluide.

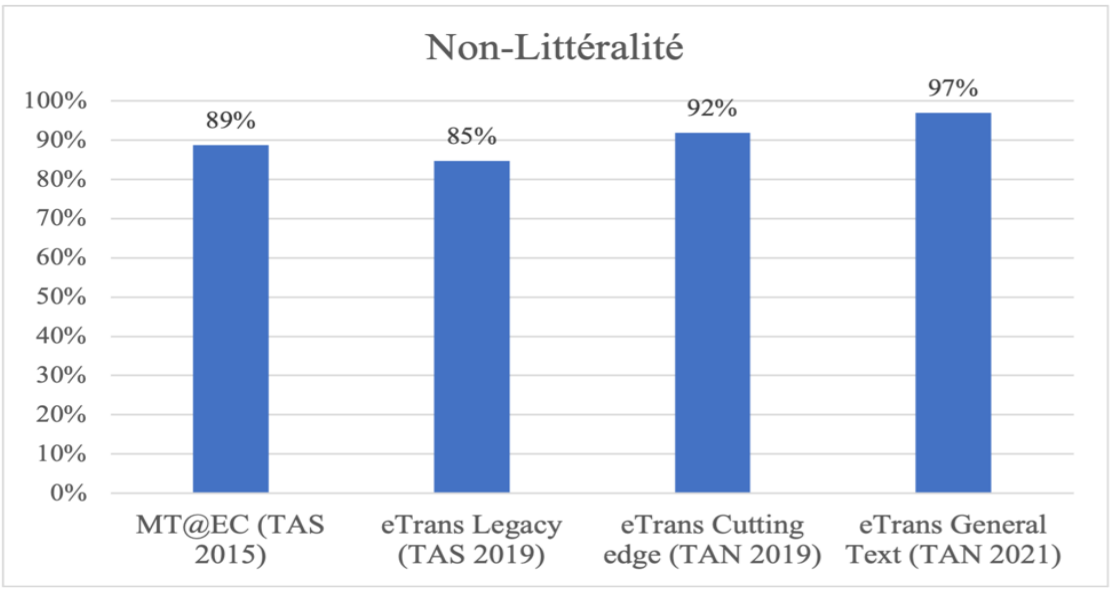
4. Resultats et Discussion

Les résultats sont présentés et discutés selon les critères définis : non-littéralité, acceptabilité, gestion de la scission et solutions originales.

4.1 Non-littéralité

Le premier graphique rend compte du pourcentage de traduction de l'UCP *cada vez* COMP sans *chaque* :

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



Graphique 1. Pourcentage de solutions non littérales

Le taux d'identification de l'UCP portugaise devant être traduite en bloc est relativement élevé. On remarque une légère baisse de la TAS de 2015 à 2019. Cela s'explique sans doute par le fait que tous les efforts d'amélioration se sont concentrés sur la TAN aux dépens de la TAS à partir de 2017. Et ces efforts semblent avoir porté leurs fruits puisqu'on constate un bond de 5% de 2019 à 2021.

L'exemple (10) illustre la supériorité de eTranslation en 2021 en termes de non-littéralité :

(10) *E, em todos os meios moscovitas, faz-se ouvir* ***cada vez mais*** *uma única pergunta:* ... (CTP)
Et dans tous les milieux moscovitas, se fait entendre* ***chaque fois plus qu'**une seule question :* ... (TAS 2015 et 2019)
En tout point de vue, une seule question est entendue* ***à chaque moment* *:* ... (TAN 2019)
Et, dans tous les médias Muscovite, une question ***de plus en plus*** *est entendue :* ... (TAN 2021)

La non-littéralité ne signifie pas que la TA soit fluide, mais les versions antérieures montrent que l'ambiguïté avec les UCP exprimant l'itération constituait encore parfois un obstacle à l'identification de l'UCP de progression et que la TAS conserve les mots de la langue source si aucun équivalent n'est disponible dans ses données.

L'exemple suivant explique le résultat plus médiocre de la TAS en 2019 :

(11) *... essas vozes cantam* ***cada vez com mais força.*** (CTP)
... ces voix chantent ***avec de plus en plus de force.*** (TAS 2015)
* *... ces voix chantent* ***chaque fois avec le plus de force.*** (TAS 2019)
... ces voix s'orientent ***de plus en plus vers une plus grande force.*** (TAN 2019)
... ces voix chantent ***de plus en plus fortement.*** (TAN 2021)

Il confirme aussi la supériorité du système de 2021 en termes d'acceptabilité, même si *fort* eût été préférable.

Un dernier exemple illustre la supériorité de la TAN selon ce premier critère :

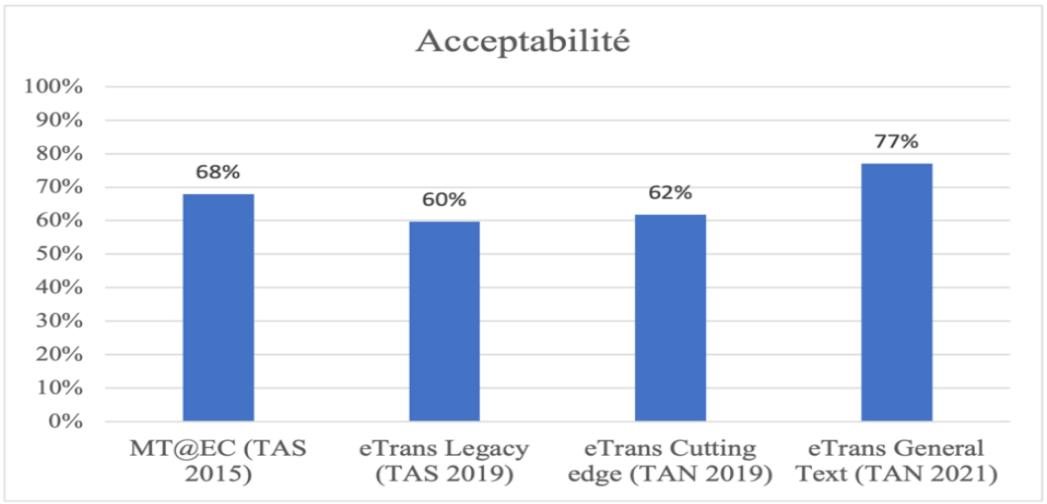
(12) ***Cada vez*** *me convenço* ***mais*** *de que em cada esquina da História assoma um nariz de Cleópatra.* (CTP)
*****Chaque fois*** *me convenço* ***plus*** *que dans chaque coin /à chaque tournant de l'histoire assoma un nez de cleópatra.* (TAS 2015 et 2019)
Je suis ***de plus en plus*** *convaincu que, dans tous les domaines de l'histoire, un nez clora.* (TAN 2019)
Je suis ***de plus en plus*** *convaincu que dans tous les coins de l'histoire il a un nez de Cléopâtre.* (TAN 2021)

L'évolution de la TA apparaît ici clairement. La traduction de la TAS est beaucoup plus littérale et la scission par un 2-gramme (*me convenço*) semble plus gênante que par un 1-gramme (*com* en 11). En effet, si la TAS produit une traduction plus littérale, quitte à conserver dans la langue cible des mots de la langue source, la TAN calcule le mot suivant le plus probable, quitte à s'éloigner du sens du texte source et à rendre la traduction inacceptable.

4.2 Acceptabilité

Les performances des systèmes de TA diminuent lorsque l'on passe du premier au deuxième critère, celui de l'acceptabilité, comme le montre le deuxième graphique :

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



Graphique 2. Pourcentage de solutions acceptables

Ici, la TAS de 2015 obtient un meilleur score que la TAS et la TAN de 2019. L'abandon de l'entretien d'*eTranslation Legacy* semble se confirmer et *eTranslation Cutting edge* en était encore à ses débuts. D'ailleurs, le pourcentage de solutions acceptables du système de TAN est nettement supérieur en 2021, même s'il reste inférieur au score obtenu selon le premier critère. Les exemples (13) à (16) apportent un peu de lumière sur ces résultats.

(13) *... para que as pessoas possam* ***trabalhar cada vez melhor***. (CTP)
... pour que les personnes puissent* ***travailler de plus en plus*. (TAS 2015 et 2019)
... pour que les citoyens puissent* ***travailler de plus en plus mieux*. (TAN 2019)
... pour que les gens puissent* ***travailler mieux et mieux*. (TAN 2021)

Ces quatre propositions de TA brute sont inacceptables, bien que la traduction ne soit pas littérale. La TAS produit un faux sens en sélectionnant l'équivalent statistiquement le plus fréquent tandis que la TAN produit un non-sens en calquant deux UCP acceptables en anglais : *increasingly better* et *better and better*. De fait, l'anglais sert souvent de langue pivot au sein des services de traduction des institutions européennes et les volumes de données dans les paires de langues comportant l'anglais sont plus importants, ce qui fait que même la TAN peut passer par l'anglais pour traduire d'une langue romane à une autre (M. Foti, communication personnelle, 1er octobre 2021).

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

L'exemple (14) comporte un jeu de mots :

(14) *Há quem, com um certo humor, defina como especialista aquele que* ***sabe cada vez mais de cada vez menos***. (CTP)
D'aucuns, avec un humour, définisse comme spécialiste celui qui* ***sait de plus en plus de moins en moins*. (TAS 2015)
D'aucuns, avec humour, de définir un certain comme celui qui* ***sait toujours plus de moins en moins*. (TAS 2019)
Certains, avec certains humours, sont définis comme des spécialistes, dont* ***on sait qu'ils sont de plus en plus connus*. (TAN 2019)
Certaines personnes, d'une certaine humeur, se définissent comme une experte qui* ***connaît de plus en plus de moins en moins*. (TAN 2021)

Ce jeu de mots constitue un fameux défi qu'aucun des systèmes de l'UE n'a été en mesure de relever, et cela ne concerne pas que les deux UCP portugaises dans cette phrase teintée d'humour. Les deux propositions de la TAS et celle de la TAN en 2021 donnent lieu à un non-sens tandis que celle de la TAN en 2019 omet de traduire une des deux UCP successives et résulte dans un faux sens avec l'ajout de *connus*.

Certes, le défi est de taille, il a pourtant été relevé par DeepL, qui a produit une traduction parfaitement idiomatique en septembre 2021 : *Il y a ceux qui, avec un certain humour, définissent le spécialiste comme celui qui* ***en sait de plus en plus sur de moins en moins de choses*** (Bacquelaine, sous presse). En effet, les données des systèmes de l'UE ne couvrent pas aussi bien la langue générale que DeepL, qui dispose d'un volume considérable de données de qualité, ce qui lui permet de surpasser ses concurrents dans la langue générale.

L'exemple suivant comporte aussi le V *saber* et *cada vez menos* y modifie trois V :

(15) *Saramago […] afirma que existe uma analfabetização lenta, que vai minando a área dos alfabetizados, que* ***sabem cada vez menos ler, escrever e «sobretudo pensar».*** (CTP)
… une analfabetização lente, qui va nuisent à la zone des alfabetizados, qui savent ***de moins en moins*** **de* ***lire, écrire et « surtout penser »***. (TAS 2015)
… une analfabetização lente, qui nuisent à la zone des instruits,* ***plus ils savent lire, écrire et « surtout penser »*. (TAS 2019)
… une maîtrise lente de l'alphabétisation, qui nuit à l'alphabétisation, qui va* ***de plus en plus souvent en savoir plus, écrire et « particulièrement penser »*. (TAN 2019)
… une lente analphabétisme, qui sape le domaine des littérés, qui ***savent de moins en moins lire, écrire et « penser particulièrement ».*** (TAN 2021)

On remarque la supériorité de la TAS en 2015 (la suppression de la préposition *de*

demande un effort minime de post-édition) et de la TAN en 2021, le contresens (proche du non-sens) produit par la TAS en 2019 et l'accumulation de défauts de la TAN en 2019. Non seulement, elle produit un contresens (*menos* devient *plus*) qui résulte dans un non-sens, avec l'ajout de *souvent*, mais aussi un calque de l'anglais (*increasingly … more* / *de plus en plus … plus*). Certes, l'ADV de quantification de fréquence *cada vez mais* (Leal, 2012) se traduit parfois par *de plus en plus souvent*, mais il ne s'agit pas de cela ici. Notons aussi l'ajout du pronom *en*, qui ne convient pas dans ce cas. Cela signifie donc que la structure *en savoir plus* fait partie de ses données et aurait donc pu être produite en (14). Outre ces ajouts, le V *ler* a été omis dans la TA.

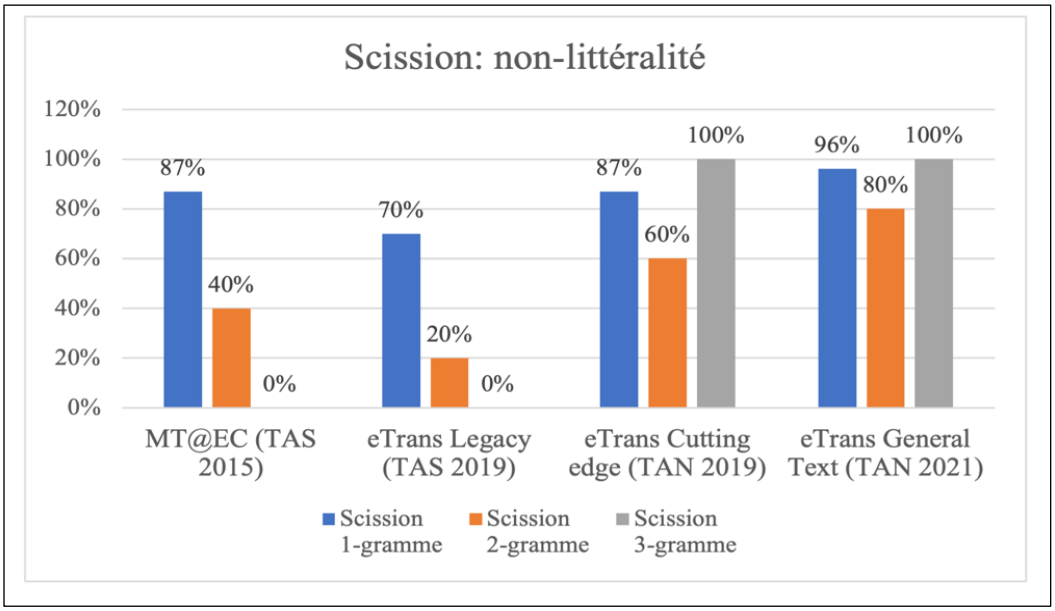
Le dernier exemple montre la supériorité de la TAN 2021 en termes d'acceptabilité :

(16) … *é importante que* ***saibamos cada vez mais da China*** … (CTP)
*… *il est important que* ***notre connaissance de plus en plus de la Chine*** … (TAS 2015 et 2019)
*… *Il est important que* ***nous volons de plus en plus de Chine*** … (TAN 2019)
… *Il est important que* ***nous en sachions de plus en plus sur la Chine*** … (TAN 2021)

Seule la solution de TAN 2021 est acceptable, toutes les autres donnent lieu à un non-sens. Comme DeepL en (14), eTranslation a produit une traduction fluide en ajoutant le pronom *en* et en utilisant la préposition *sur* au lieu de *de*. C'est sans doute la succession inhabituelle de deux UCP dans le jeu de mots en (14) qui a compromis l'acceptabilité de la TAN 2021.

4.3 Gestion de la scission

Sur les 29 UCP scindées, 23 le sont par un seul mot, cinq par deux mots et une seule par trois mots. En ce qui concerne la scission de 3-gramme, une seule occurrence empêche toute généralisation, mais on constate quand même que plus la distance entre l'élément ordonnant *cada vez* et COMP augmente, plus l'identification de l'UCP à traduire en bloc se complique, surtout pour la TAS. Commençons par la non-littéralité présentée dans le troisième graphique :



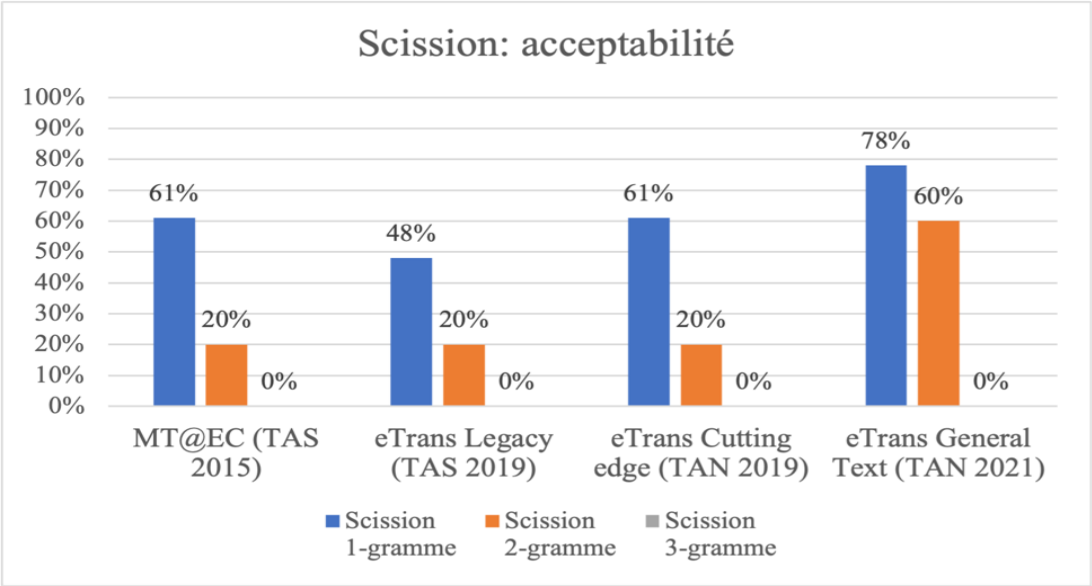
Graphique 3. Pourcentage de non-littéralité en cas de scission de l'UCP

Les scores sont plus faibles et suggèrent que la scission constitue effectivement un obstacle à l'identification de l'UCP à traduire en bloc. Néanmoins, la TAN a identifié l'UCP scindée par trois mots et TAN 2021 obtient le meilleur score dans tous les cas. La TAN 2019 s'en sort mieux que la TAS en cas de scission par deux mots, mais elle obtient le même score que la TAS 2015 en cas de scission par un seul mot. La lanterne rouge est à nouveau la TAS 2019. Ces résultats confirment la nécessité d'une intervention humaine régulière pour corriger les défauts d'apprentissage automatique, profond (TAN) ou non (TAS).

Les exemples (11) et (12) illustrent déjà ces résultats pour les scissions de 1-gramme et de 2-gramme. L'exemple (17) témoigne de la supériorité de TAN 2021 quant à la non-littéralité :

(17) *E **cada vez** chega **menos água do Douro** até ao oceano Atlântico*. (CTP)
Et **chaque fois arrive **moins d'eau du Douro/Duero** jusqu'à l'océan Atlantique*. (TAS 2015 et 2019)
Et **chaque fois que vous arriverez **moins d'eau du Douro** à l'océan Atlantique.* (TAN 2019)
*Et **de moins en moins d'eau** vient **du Douro** à l'océan Atlantique.* (TAN 2021)

Pour ce qui est de la scission de 3-gramme, voyons d'abord les résultats en termes d'acceptabilité :



Graphique 4. Pourcentage d'acceptabilité en cas de scission de l'UCP

Ici aussi, TAN 2021 domine, mais TAS 2015 et TAN 2019 sont au coude à coude. Cependant, les TA acceptables ne concernent pas nécessairement les mêmes occurrences de l'UCP portugaise, comme le montrent les exemples (11) et (12). En (11), la solution de TAS 2015 est acceptable et en (12), c'est celle de TAN 2019. Voici un exemple où les deux solutions sont acceptables, comme celle de TAN 2021, mais contrairement à celle de TAS 2019, littérale et donc inacceptable :

(18) *Mas **cada vez há mais factores** que fogem ao controlo dos sindicatos.* (CTP)
*Mais **il existe de plus en plus de facteurs** qui échappent au contrôle des syndicats.* (TAS 2015)
Mais **chaque fois il y a plusieurs facteurs qui échappent au contrôle des syndicats.* (TAS 2019)
*Mais **il y a de plus en plus de facteurs** qui fuient le contrôle des syndicats.* (TAN 2019)
*Mais **il y a de plus en plus de facteurs** qui échappent au contrôle des syndicats.* (TAN 2021)

Dans l'exemple (19), il s'agit d'une phrase simple. Le défi provient sans doute du fait que le 2-gramme qui scinde l'UCP est un V pronominal, que les Portugais désignent par le terme 'voix passive synthétique ou pronominale', dont l'équivalent le plus fréquent en français passe par le pronom *on*.

(19) ***Cada vez*** *se lê* ***menos***. (CTP)
*****Chaque fois*** *lieu* ***de moins***. (TAS 2015 et 2019)
*****Pour le moment,*** *il est* ***de plus en plus facile*** *de lire.* (TAN 2019)
*****Chaque fois que*** *vous lisez* ***moins***. (TAN 2021)

TAN 2019 traduit sans *chaque*, mais un contresens (*de plus en plus*) s'ajoute à un faux sens (*Pour le moment*) et on se demande d'où sort l'ADJ *facile*, si ce n'est le mot le plus probable indépendamment du texte source. Tous les autres produisent une traduction littérale inacceptable et la voix passive synthétique s'avère problématique pour tous les systèmes.

Le dernier exemple concerne la scission par un 3-gramme :

(20) ***Cada vez*** *as regras imperam* ***menos,*** *cada vez mais tudo é permitido*. (CTP)
****Chaque fois*** *les règles appliquées* ***moins*** *de plus en plus tout est permis.* (TAS 2015*)*
*****Chaque fois*** *les règles* ***moins*** *prises en compte, de plus en plus, tout est permis.* (TAS 2019)
Les règles sont* ***de moins en moins nombreuses* *et de plus en plus autorisées.* (TAN 2019)
*****De plus en plus****, les règles prévalent* ***de moins en moins****, de plus en plus tout est autorisé.* (TAN 2021)

La TAS propose une TA littérale qui n'a pas de sens. La TAN produit une TA sans *chaque* en 2019 comme en 2021, mais ces TA sont inacceptables. En 2019, il s'agit d'un faux sens et en 2021, le calque sur l'anglais résulte de nouveau dans un non-sens.

4.4 Solutions originales

Rappelons que les solutions originales acceptables ont été classées dans trois catégories : la lexicalisation, le changement de catégorie grammaticale ou recatégorisation et la naturalisation.

Les solutions originales dignes d'un biotraducteur sont rares, mais tous les systèmes en ont produit, même si elles sont plus diversifiées en TAN qu'en TAS. TAS 2015 produit six recatégorisations. TAS 2019 en produit quatre et une lexicalisation. TAN 2019 propose trois lexicalisations, trois recatégorisations et quatre naturalisations, et nous avons relevé quatre lexicalisations, deux recatégorisations et cinq naturalisations dans la TA brute de TAN 2021.

4.4.1 Lexicalisation

Dans l'exemple (21), la lexicalisation par un V de TAN 2019 est acceptable, mais TAN 2021 calque un anglicisme qui la rend inacceptable :

(21) *A Alemanha* ***está a ficar cada vez mais velha****.* (CTP)

L'Allemagne ***vieillit.*** (TAN 2019)
L'Allemagne* ***vieillit et vieillit.* (TAN 2021)

L'exemple (22) présente la seule lexicalisation de TAS 2019, où l'UCP portugaise a été rendue par un ADJ verbal :

(22) *..., o «buraco» está* ***cada vez maior****, ...* (CTP)
..., le déficit est ***croissant****, ...* (TAS 2019)

Pour compléter cette série, l'exemple (23) propose une lexicalisation réussie par la TAN :

(23) *Segundo Paula Nobre,* ***cada vez há maior procura*** *deste tipo de passeios.* (CTP)
Selon Paula Nobre, il existe une ***demande croissante*** *pour ce type de trottoirs.* (TAN 2019)
Selon Paula Nobre, il y a une ***demande croissante*** *pour ce type de visites.* (TAN 2021)

La TAN produit la même lexicalisation en 2019 et en 2021 alors que la TAS 2019 propose la solution la plus fréquente et que la TAS 2015 produit une traduction littérale inacceptable. Notons que l'ambiguïté lexicale de *passeio* (*trottoir*, *promenade*, *excursion*, *visite*) n'a pu être levée que par TAN 2021. Certes, l'absence de contexte ne facilite pas les choses.

4.4.2 Recatégorisation

La recatégorisation d'un N en ADJ a été proposée par tous les systèmes, sauf TAN 2019 dont la TA donne lieu à un non-sens dans l'exemple (24) :

(24) *Os que não trabalham, e que* ***são cada vez em maior número****,* … (CTP)
Ceux qui sont sans emploi/qui ne travaillent pas, et qui ***sont de plus en plus nombreux****,* … (TAS 2015, TAS 2019, TAN 2021)
Ceux qui ne travaillent pas et qui,* ***de plus en plus, ne travaillent pas**,* … (TAN 2019)

En (25), la TAS recatégorise un ADV en ADJ alors que la TAN conserve l'ADV:

(25) *... acontece* ***cada vez mais raramente****.* (CTP)
... c'est ***de plus en plus rare****.* (TAS 2015 et 2019)

La traduction du V *acontecer* par *être* implique l'emploi d'un ADJ (attribut du sujet) et cet exemple aurait pu être classé parmi les naturalisations, mais tout classement implique des choix discutables. La solution de TAN 2019 (*être de plus en plus rarement*

le cas) a été considérée comme une naturalisation, tandis que *se passer de plus en plus rarement* (TAN 2021) a été considéré comme acceptable, mais peu idiomatique.

4.4.3 Naturalisation

Dans notre échantillon de TA brute, seuls les systèmes de TAN ont produit des solutions témoignant d'une naturalisation digne d'un biotraducteur. On vient d'en voir un exemple, en voici deux autres :

(26) *... o «multimedia» e a integração de todas as formas de intercâmbio de informação* ***terão um peso cada vez maior***. (CTP)
... le « multimédia » et l'intégration de toutes les formes d'échange d'informations ***joueront un rôle de plus en plus important***. (TAN 2019)

(27) ***Cada vez*** *se torna* ***mais indispensável e insubstituível*** *uma política de verdade, uma política descentralizada, uma política adequada ao país real.* (CTP)
Une véritable politique, une politique décentralisée, une politique adaptée au vrai pays devient ***de plus en plus indispensable et irremplaçable***. (TAN 2021)

En (26), la naturalisation résulte d'une équivalence phraséologique entre *ter um peso grande* et *jouer un rôle important*. Tous les systèmes ont proposé une solution acceptable, mais celle de TAN 2019 nous semble plus fluide que celles de la TAS (*avoir une grande incidence* en 2015 et *avoir une incidence importante* en 2019), même si *peso* a été traduit par *incidence*, ou celle de TAN 2021 qui passe par une lexicalisation de l'UCP portugaise, mais calque *ter um peso* : *auront un poids croissant*.

Par contre, l'exemple (27) témoigne de la capacité de restructuration profonde de TAN 2021, alors que TAN 2019 n'en semble pas capable.

Dans notre échantillon de TA brute, la recatégorisation apparaît ainsi comme la solution originale la plus accessible à la machine, dès la génération de TAS. La lexicalisation s'avère plus compliquée avant l'avènement de l'apprentissage profond et la naturalisation est l'apanage de la TAN. Cependant, certaines solutions qui auraient pu être originales donnent lieu à des solutions inacceptables. On l'a vu dans l'exemple (21), où l'anglicisme rend la lexicalisation inacceptable.

5. Conclusion : de la TAS à la TAN

Dans l'ensemble, les résultats de cette étude de la TA brute en termes de non-littéralité, d'acceptabilité et de solutions originales confirment le constat de Koehn (2021) : la TA s'améliore et devient plus utile, mais c'est toujours un outil imparfait. Malgré les progrès constants de l'intelligence artificielle, la machine peine encore à lever l'ambiguïté lexicale, phraséologique ou syntaxique, car elle est dépourvue de bon sens et de connaissance du monde, et les données sont toujours limitées et ne peuvent répondre à tous les besoins (Koehn, 2020, pp. 5-8).

Le déclin des performances de la TAS entre les résultats de 2015 et ceux de 2019 démontre l'importance de l'entretien des systèmes par des humains. Étant donné que les systèmes 'apprennent' automatiquement, il faut que les humains s'efforcent inlassablement de corriger les erreurs et d'enrichir les données de biotraduction de qualité sur lesquelles se fondent ces systèmes (Foti, 2021) pour atteindre un équilibre optimal entre fluidité et adéquation.

Les systèmes de TAS sont entraînés sur des corpus alignés bilingues et nous n'avons observé aucun anglicisme dans notre échantillon. Par contre, eTranslation tend à utiliser l'anglais comme langue pivot quelle que soit la paire de langues (Foti, communication personnelle, 1er octobre 2021). Aux yeux des informaticiens, qui raisonnent le plus souvent en termes quantitatifs, le passage par l'anglais se justifie sans doute par le volume plus important de mémoires de traduction dans les paires EN-FR et EN-PT. Aux yeux du linguiste, ce détour semble toutefois farfelu voire dangereux.

La mondialisation et l'intensification des échanges qu'elle engendre provoque une augmentation des besoins en traduction. Il suffit de penser au portail Re-open UE, dont la mise à jour par des biotraducteurs s'est avérée impossible en raison du rythme effréné des changements de mesures sanitaires dans les différents États membres de l'UE (Foti, 2021). La TA fait désormais partie du poste de travail du biotraducteur, qui aurait d'ailleurs tort de s'en passer. Ainsi, la formation des futurs traducteurs doit les préparer à comprendre et à suivre l'évolution de la TAN pour pouvoir pratiquer la post-édition en connaissance de cause, sans se laisser tromper par l'apparente fluidité qui cache encore bien souvent des contresens et des faux sens.

Références

[1] Bacquelaine, F. (sous presse). DeepL et Google Translate face à l'ambiguïté phraséologique. *Journal of Data Mining and Digital Humanities, Vers une robotique du traduire*.

[2] Bacquelaine, F. (2020). *Traduction humaine et traduction automatique du quantificateur universel portugais 'cada' en français et en anglais. Étude de phraséologie comparée* [Thèse de doctorat, Université de Porto]. Repositório Aberto da Universidade do Porto. https://hdl.handle.net/10216/127715

[3] Ballard, M. (2006). À propos des procédés de traduction. *Palimpsestes, Hors série | 2006*, 113-130. https://doi.org/10.4000/palimpsestes.236

[4] CETEM Público: Rocha, P. & Santos, D. (2000). CETEM Público: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In M. das G. Volpe Nunes (Ed.), Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000) (Atibaia, São Paulo, Brasil, 19 a 22 de Novembro de 2000), (pp. 131-140).

[5] Durieux, C. (2014). L'unité de traduction : une unité de sens. In S. Mejri, I. Sfar, & M. Van Campenhoudt (dir.), *L'unité en sciences du langage. Actes des Neuvièmes journées scientifiques du réseau thématique Lexicologie,*

terminologie, traduction, Paris, 15 et 16 septembre 2011 (pp. 381-388). Éditions des archives contemporaines.

[6] Eisele, A. (2017). From MT@EC to eTranslation in CEF. Overview for the DGT QT21 Workshop (Luxembourg, 17 mars 2017) [Diaporama]. http://www.qt21.eu/wp-content/uploads/2018/02/02-20170315-DGT-overview-March-2017.pdf

[7] Europarl V7: Koehn, P. (2005). A Parallel Corpus for Statistical Machine Translation. In Proceedings of Machine Translation Summit X: Papers (pp. 79–86). Association for Computational Linguistics.

[8] Foti, M. (1/10/2021). eTranslation et l'Europe numérique : Technologies et données linguistiques [Podcast]. POD Unistra. https://pod.unistra.fr/video/45872-robotrad-etranslation-et-leurope-numerique-technologies-et-donnees-linguistiques/

[9] Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), Phraseology: An Interdisciplinary Perspective (pp. 27-49). John Benjamins Publishing Company. https://doi.org/10.1075/z.1 39

[10] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T, Hoang, H., Heafield, K, Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., & Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. In Proceedings of ACL 2018, System Demonstrations (pp. 116–121). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-4020

[11] Koehn, P. (2021, 17 juin). Applying New Advances in AI-based Machine Translation to Real World Use [Webinaire]. https://omniscien.com/webinars/applying-new-advances-in-ai-based-machine-translation-to-real-world-use-cases/

[12] Koehn, P. (2020). Neural Machine Translation. Cambridge University Press. https://doi.org/10.1017/9781108608480

[13] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In S. Ananiadou (Ed), Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions (pp. 177–180). Association for Computational Linguistics.

[14] Leal, A. (2012). Cada vez mais/menos : comparative construction or quantification over eventualities ? In C. Schnedecker & C. Armbrecht (Éd.), La quantification et ses domaines : actes du colloque de Strasbourg 19-21 octobre 2006 (pp. 355-366). Honoré Champion.

[15] Nádvorníková, O. (2016). Le corpus multilingue InterCorp et les possibilités de son exploitation. In E. Buchi, J.-P. Chauveau, & J.-M. Pierrel (éd.), Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Vol.2) (pp. 16351649). ÉLiPhi.

[16] Nothomb, A. (1999). *Stupeur et tremblements*. Albin Michel.
[17] Nothomb, A. (1999), traduction de Carlos de Almeida (2000). *Temor e Tremor*. Bizâncio.
[18] Schmale, G. (2013). Qu'est-ce qui est préfabriqué dans la langue ? – Réflexions au sujet d'une définition élargie de la préformation langagière. *Langages*, 189, 27-45.
[19] Theissen, A. (2007). Quantification universelle : chaque fois / toutes les fois. *Verbum*, XXIX (3-4), 243-257.
[20] Tiedermann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.). Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012) (pp. 2215-2218). European Language Resources Association.

Remerciements
Cette recherche a été financée par des fonds publics portugais et des fonds communautaires européens alloués par la Fondation pour la Science et la Technologie (FCT, Portugal) au Centre de Linguistique de l'Université de Porto dans le cadre du programme de financement FCT-UIDB/00022/2020.
Mes remerciements vont également au réseau LTT, au Ministère de la Culture français et à l'Università degli Studi di Napoli « L'Orientale », qui ont permis la réalisation du colloque international « La traduction au service des institutions : outils, expérimentations et innovations pour le multilinguisme » les 4 et 5 novembre 2021, ainsi qu'à l'équipe du Centre de Recherche sur l'Information scientifique et technique (CERIST), qui héberge la revue TRANSLANG sur ASJP.

Notice biographique de l'Auteur
Françoise Bacquelaine est née en Belgique, où elle a obtenu une licence en Philologie germanique à l'Université de Liège. Titulaire d'un diplôme d'université de traducteur généraliste délivré par l'Université de Haute Bretagne, Rennes 2, elle a soutenu un mémoire de master en terminologie et une thèse de doctorat en traduction à l'Université de Porto (Portugal). Elle a enseigné l'anglais et l'allemand en Belgique et au Maroc avant de s'installer au Portugal, où elle a été lectrice de français et est actuellement maître de conférences à l'Université de Porto. Elle y enseigne essentiellement la traduction générale et spécialisée entre le français et le portugais. Ses recherches actuelles portent sur la terminologie, la didactique de la traduction, la traduction automatique, la post-édition et la quantification universelle.