DeepL et Google Translate face à l'ambiguïté phraséologique

Françoise Bacquelaine

DeepL et Google Translate face à l'ambiguïté phraséologique

Françoise Bacquelaine[1*]
1 Université de Porto, Portugal[1]
*franba@letras.up.pt

Résumé
Malgré les progrès de la traduction automatique neuronale, l'intelligence artificielle ne permet toujours pas à la machine de comprendre pour déjouer tous les pièges de la traduction, notamment ceux de l'ambiguïté lexicale, phraséologique, syntaxique et sémantique (Koehn 2020). Deux structures portugaises moyennement figées présentent les caractéristiques des « unités de construction préformées » (UCP) décrites par Schmale (2013). Elles relèvent donc de la phraséologie au sens large et doivent être traduites en bloc. Les principaux défis de traduction en bloc que lancent ces UCP binaires à la machine résultent, d'une part, de variables simples ou complexes, et, d'autre part, des propriétés syntaxiques de scission et d'inversion des éléments sur l'axe syntagmatique. Un échantillon de 168 occurrences de ces UCP en contexte phrastique a été prélevé sur un corpus journalistique portugais. Cet échantillon a été traduit en français par DeepL et Google Translate en 2019 et en 2021. Les traductions automatiques brutes ont été confrontées à un modèle de biotraduction établi à partir de corpus parallèles ou alignés portugais-français et analysées en fonction de deux critères généraux (non-littéralité et acceptabilité) et de quelques défis spécifiques à chaque UCP. Cette analyse permet d'évaluer l'évolution de ces deux systèmes de traduction automatique face à l'ambiguïté phraséologique et d'en tirer des conclusions quant à la possibilité d'extinction de la biotraduction et aux implications de ces outils performants sur la formation des futurs prestataires de services linguistiques.

Mots-clefs
traduction automatique neuronale ; post-édition ; levée d'ambiguïté ; unité de construction préformée ; portugais ; français

INTRODUCTION
En 2009, Yorick Wilks critique à la fois l'engouement excessif de certains pour la traduction automatique (TA) statistique et le scepticisme d'autres face à la robotisation du traduire : « Machine translation is not, as some believe, solved, nor is it impossible, as others still claim » (Wilks 2009 : v). Depuis le milieu des années 2010, la TA est devenue neuronale et a connu un bond qualitatif indéniable, mais tous les problèmes ne sont pas résolus : « The reality is that MT has become better over time and it has become more useful for more use cases. […] So, we are actually at the point where it is quite useful, but it is definitely not solved » (Koehn 2021). Certes, la TA donne d'excellents résultats dans des domaines bien délimités, lorsque les données sont de bonne qualité et en quantité suffisante, mais il s'agit ici de tester deux systèmes de TA accessibles au grand public sur un échantillon de langue générale, où les possibilités du langage sont infinies (Melby et Warner 1995).
Cet article se divise en trois parties. Il s'agit d'abord de décrire les unités phraséologiques portugaises à traduire en bloc et les défis qu'elles lancent tant au traducteur humain qu'à la

[1] Recherche financée par des fonds portugais et communautaires européens octroyés par la *Fundação para a Ciência e a Tecnologia* (Portugal) au *Centro de Linguística da Universidade do Porto* dans le cadre du programme de financement FCT-UIDB/00022/2020.

machine. La deuxième partie est consacrée à la méthodologie. Y sont présentés les corpus, les outils, l'échantillon de test et l'échantillon testé, le modèle de traduction humaine et les critères d'anlyse de la TA. Les résultats du test sont commentés dans la troisième partie.

I DEUX UNITES DE CONSTRUCTION PREFORMEES ET TROIS UNITES DE TRADUCTION

Il est généralement admis que le concept d'unité en phraséologie et en traduction est un concept à géométrie variable. Schmale (2013) propose une « définition élargie de la préformation langagière » et le terme général d'*unité de construction préformée* (UCP), qui peut s'appliquer à deux unités phraséologiques portugaises exprimant la progression (*cada vez* COMP) et la proportion (SQ_1 PREP *cada* SQ_2). En effet, il s'agit d'unités polylexicales (« polyfactorielles » pour Schmale, qui élargit le concept au non-verbal) présentant un certain degré de « stabilité ou de figement structural/lexical » (*ibidem* : 42) qui « permet à un membre de la communauté langagière concernée de [les] reconnaître et de [les] réutiliser » (*ibidem* : 41). Leur sens est relativement compositionnel, mais leur traduction mot à mot en français est impossible, ce qui leur confère un certain degré d'idiomaticité (*ibidem* : 35). Elles ne présentent pas de saturation lexicale et les éléments qui les composent ne sont pas nécessairement contigus. Ces deux UCP correspondent à trois unités de sens (progression, proportion entre un ensemble et un sous-ensemble et proportion entre deux ensembles) et donc trois unités de traduction (UT), puisque l'unité de sens conditionne l'UT : « Tant qu'il n'y a pas d'unité de sens construite, il ne peut y avoir d'unité de traduction, l'une se superposant à l'autre » (Durieux 2014 : 386).
Nous présentons les trois unités à traduire en bloc et les défis particuliers qu'elles recèlent.

1.1 La progression et la proportion

La première UCP portugaise exprimant la progression est particulièrement fréquente et disséminée à travers la plupart des registres. La deuxième l'est beaucoup moins, mais partage avec la première l'emploi idiomatique du quantificateur universel *cada*, puisque *chaque* est impossible pour les traduire en français.

1.1.1 La progression : cada vez *COMP*

En portugais, la progression quantitative ou qualitative s'exprime le plus souvent au moyen de l'élément ordonnant *cada vez* suivi d'un comparatif simple (*mais, menos, melhor, pior, maior, menor*) ou complexe, lorsque *mais* ou *menos* modifie un adjectif (ADJ), un adverbe (ADV) ou un nom (N), comme dans le premier exemple :

Exemple 1 (Europarl ; langue source (LS) : français)
… a delirante administração comunitária e os danos que ela provoca só podem […] dar origem a <u>cada vez mais recursos</u>.
… la délirante administration communautaire et les méfaits qu'elle engendre ne peuvent […qu'] engendrer des <u>recours de plus en plus nombreux</u>.

On remarque non seulement que la traduction en français ne comporte pas *chaque*, mais aussi que le comparatif *mais* opère sur un N tandis que *de plus en plus* opère sur un ADJ. L'exemple 2 illustre l'inverse, *maior* est le comparatif synthétique de l'ADJ *grande* tandis que *de plus en plus* quantifie un N :

Exemple 2 (A. Nothomb)
Virou as páginas com um frenesim cada vez maior.
Il tourna les pages avec de plus en plus de frénésie.

1.1.2 La proportion : SQ_1 PREP cada *SQ_2*
L'UCP SQ_1 PREP *cada* SQ_2 exprime soit la proportion entre un ensemble et un sous-ensemble, soit la proportion entre deux ensembles. Dans le premier cas, PREP est le plus souvent *em* et dans le deuxième *por*. Les exemples 3 et 4 illustrent ces PREP les plus fréquentes et leur équivalent habituel en français (*sur* et *pour*).

Exemple 3 (Europarl ; LS : suédois)
Imaginemos que um em cada dois homens fosse objecto de assédio sexual.
Imaginez qu'un homme sur deux soit exposé au harcèlement sexuel.

Notons que N (*homens*, *homme*) précède PREP en français alors qu'il la suit en portugais, où N est précédé d'un numéral (NUM) cardinal (*dois*) avec lequel il forme un individu collectif que *cada* peut modifier (Leal 2006), ce qui est impossible avec *chaque*. Ce trait combinatoire distinctif entre *cada* et *chaque* implique donc une traduction en bloc de cette UCP, que la proportion soit établie entre un ensemble et un sous-ensemble (Ex. 3) ou entre deux ensembles (Ex.4).

Exemple 4 ((Europarl ; LS : néerlandais)
Diz-se que, na Flandres, existe um computador por cada dez alunos.
On dit qu'en Flandre, il y a un ordinateur pour dix élèves.

Les deux UCP portugaises correspondent ainsi à trois UT. L'identification de ces UT complexes permettant de les traduire en bloc se heurte à plusieurs défis.

1.2 Défis particuliers de ces unités de traduction
Les défis particuliers de ces trois UT sont d'ordre lexical, sémantique et syntaxique. Elles ne sont pas saturées lexicalement puisqu'elles contiennent une (COMP) ou plusieurs variables (PREP et deux SQ). Leur degré de figement lexical est donc moyen dans le cas de la progression et très faible dans le cas de la proportion, ce qui peut entraîner une ambiguïté phraséologique. Enfin, au niveau syntaxique, l'UCP exprimant la progression admet la scission et celle exprimant la proportion admet non seulement la scission, mais aussi l'inversion.

1.2.1 Variables d'un ou plusieurs mots

Au niveau lexical, la ou les variables comportent un ou plusieurs mots. PREP n'en comporte jamais qu'un, mais COMP et SQ_1 peuvent être simples ou complexes et SQ2 se compose au moins de NUM cardinal et de N. Cette particularité se manifeste dans les exemples ci-dessus et ci-dessous.

1.2.2 Ambiguïté phraséologique

Au niveau sémantique, l'élément ordonnant de l'UCP exprimant la progression, le bigramme *cada vez*, est extrêmement fréquent en portugais d'Europe et entre dans la formation d'autres UCP, qui se traduisent avec ou sans *chaque*. L'UCP (*de*) *cada vez que*, fortement concurrencée par *sempre que*, correspond à (*à*) *chaque fois que* et l'ambiguïté est facilement levée. Mais la locution adverbiale *de cada vez* peut signifier (*à*) *chaque fois* ou *à la fois*, selon qu'il s'agit d'une simple itération ou de ce que nous avons appelé le « compte-gouttes » (Bacquelaine 2020). L'exemple 5 montre trois occurrences du trigramme *de cada vez* correspondant à des UCP et des UT différentes.

Exemple 5 (Europarl)
É necessário proceder, de cada vez, a uma nova repartição.
À chaque fois, il faut procéder à une nouvelle répartition. (LS : néerlandais)
... a Comissão pretende dar um passo de cada vez ...
... la Commission veut faire un pas à la fois … (Barnier)
Perante a existência de cada vez mais ameaças, ...
Confrontés à des menaces de plus en plus nombreuses, ... (LS : allemand)

Le seul élément constant de l'UCP exprimant la proportion, c'est le quantificateur universel *cada* pouvant se combiner à un NUM cardinal différent de l'unité mathématique. D'emblée, cette UCP est ambiguë puisqu'elle donne lieu à deux unités de sens, deux UT, dont les équivalents habituels en français diffèrent, alors que les mêmes PREP sont possibles en portugais. En outre, ces UT sont beaucoup moins fréquentes que la première et beaucoup moins disséminées à travers les divers registres, ce qui constitue un obstacle à la TA de la langue générale. L'exemple 6 montre que la PREP *por* peut exprimer une proportion entre un ensemble et un sous-ensemble, que la PREP *em* peut exprimer une proportion entre un ensemble d'unités de temps et un ensemble d'événements périodiques, et que d'autres prépositions sont aussi possibles pour exprimer des proportions.

Exemple 6 (Europarl)
De facto, existem 50 albinos <u>por</u> cada milhão de pessoas.
En fait, 50 personnes <u>sur</u> un million sont albinos. (LS : polonais)
… <u>em</u> cada 4 segundos morre de fome uma pessoa.
... toutes les quatre secondes, un être humain meurt de faim. (LS : suédois)
Apenas uma <u>de</u> cada dez mil pessoas em África é cientista ou engenheiro, ...
Sur ce continent, seulement une personne <u>sur</u> 10 000 est scientifique ou ingénieur, ... (Fiona Hall)
Em Roma, existem sete carros <u>para</u> cada dez residentes, ...
À Rome, on compte 7 voitures <u>pour</u> 10 résidents, ... (LS : italien)

1.2.3 Défis syntaxiques : scission et inversion

Au niveau syntaxique, la scission et l'inversion peuvent également compliquer la tâche du traducteur humain ou de la machine. L'UCP exprimant la progression n'admet pas l'inversion, mais la scission est assez courante. Dans l'exemple 7, l'élément ordonnant *cada vez* est séparé de COMP par le V :

Exemple 7 (Europarl)

... cada vez se torna mais difícil explicar aos agricultores europeus o que devem ou não fazer e quais os motivos, ...
… il est de plus en plus difficile d'expliquer aux agriculteurs européens ce qu'ils peuvent ou ne peuvent pas faire pour telle ou telle raison, … (Nassauer)

Dans l'exemple 8, il s'agit de proportion entre un ensemble et un sous-ensemble. Dans la première occurrence, PREP SQ_2 précède SQ_1 (inversion) et l'inversion se combine à la scission dans la deuxième :

Exemple 8 (Europarl)
... na União Europeia, em 1996, em cada 100 mil trabalhadores, 4.229 estiveram com baixa mais de três dias devido a acidente de trabalho.
... dans l'Union européenne, 4 229 travailleurs sur 100 000 se sont arrêtés de travailler plus de trois jours à cause d'un accident du travail, en 1996. (LS : portugais)
Segundo a Amnistia Internacional, desde a instauração da pena capital nos Estados Unidos, em cada seis presos executados, há um inocente condenado à morte, posteriormente ilibado.
Selon Amnesty International, depuis la restauration de la peine capitale aux États-Unis, un prisonnier exécuté sur six est innocenté après sa condamnation à mort. (LS : espagnol)

En français, la position de N en SQ_1 rend l'inversion et la scission plus difficiles, mais elles ne sont pas impossibles pour autant. Quant à la proportion entre deux ensembles, la présence de deux N permet plus facilement l'inversion et la scission en français, comme on peut le constater dans l'exemple 9.

Exemple 9 (Europarl)
Em 1970, por cada 100 homens que exerciam uma profissão somente 37 mulheres se encontravam na mesma situação.
En 1970, pour 100 hommes occupant un emploi, on dénombrait 37 femmes dans la même situation. (Junker)

On le voit, l'identification de ces trois UT présente plusieurs défis que nous avons lancés à la TA.

II MÉTHODOLOGIE
Pour comparer les performances de la TA neuronale accessible au grand public, nous avons d'abord établi un modèle de biotraduction à partir de corpus parallèles et alignés. Nous avons ensuite sélectionné un échantillon de test représentatif de la diversité d'emplois et de défis potentiels. Il n'est donc pas représentatif de l'usage le plus courant, par exenple tous les COMP y sont réunis, alors que *mais* est de loin le plus fréquent. Les outils sont Google Translate et DeepL, deux outils bien connus du grand public. L'échantillon a été testé en août 2019 et en septembre 2021. Les critères d'analyse de la TA brute se fondent sur le modèle de biotraduction.
Cette deuxième partie se divise ainsi en quatre sous-parties. La première décrit le corpus parallèle et les corpus alignés portugais-français, ainsi que le modèle de traduction. La deuxième présente les systèmes de TA neuronale, la troisième l'échantillon de test, et la quatrième les critères d'analyse.

2.1 En quête de modèle de biotraduction

L'exploration de corpus parallèles et alignés permet de dégager un modèle de biotraduction à condition que les biotraducteurs produisent des traductions de qualité. Ce qui nous semble être le cas ici.

2.1.1 Corpus parallèles et alignés

Le modèle de biotraduction des trois UT se fonde sur quatre corpus, dont un parallèle et trois alignés. Le corpus parallèle est un corpus littéraire. Il s'agit du roman d'Amélie Nothomb *Stupeur et tremlements* (1999) et de sa traduction en portugais par Carlos de Almeida, *Temor e Tremor* (2000). La traduction de ce roman compte 111 pages et contient trente occurrences de *cada*, dont treize de *cada vez* COMP, mais aucune proportion. Deux corpus alignés ont été explorés en ligne grâce à l'interface de recherche multilingue du site de Tiedermann (2012) : *Opus multilingual search interface*. Le premier émane de l'Agence européenne des médicaments (EMEA) et contient 1 105 678 bi-segments PT-FR (presque 28 millions de mots) et 388 658 bi-segments FR-PT (presque 13 millions de mots). En raison de sa spécificité, il nous a permis notamment de mobiliser des équivalents de traduction de la proportion. Le deuxième est la septième version du corpus Europarl comportant les mémoires de traduction du Parlement européen produites entre 1996 et 2011. Il contient près de deux millions de bi-segments PT-FR (1 980 132) et FR-PT (1 941 809), ce qui représente environ 50 millions de mots par langue. Vu la nature du corpus, il s'agit généralement de discours oral, proche de la langue courante, même si les thèmes abordés sont propres à l'UE. Le troisième corpus aligné, Intercorp, résulte d'une initiative tchèque (Institut du Corpus National Tchèque de l'Université Charles à Prague). Son interface permet à l'utilisateur de choisir ses langues de travail (notamment la variante européenne du portugais) et de définir un sous-corpus en fonction de ses besoins (Nádvorníková 2016 : 1643). Notre sous-corpus se compose d'un corpus journalistique (PressEurop) aligné automatiquement, de mémoires de traduction de l'Acquis communautaire et d'Europarl, ainsi que de trois corpus littéraires alignés manuellement. Il s'agit de trois romans bien connus en anglais et de leurs traductions en français et en portugais d'Europe : *Alice in Wonderland* de Lewis Carroll, *Harry Potter and the Philosopher's Stone* de J. K. Rowling et *The Fellowship of the Ring* de J. R. R. Tolkien. Ce sous-corpus nous a fourni plusieurs centaines d'occurrences de l'UT exprimant la progression et quelques dizaines des deux types de proportion.
Ensemble, ces quatre corpus nous semblent assez diversifiés pour dégager un modèle de biotraduction de qualité.

2.1.2 Modèle de biotraduction

La plupart des équivalents figurant dans le modèle de biotraduction ci-dessous (Tableau 1) ont été révélés dans les exemples 1 à 4 et 6 à 9, ainsi que dans la dernière occurrence de l'exemple 5. L'exemple 10 illustre la traduction de la progression par *toujours* COMP et une lexicalisation de progression qualitative, alors que l'exemple 1 peut être considéré comme une lexicalisation quantitative (*cada vez mais recursos* / *des recours de plus en plus nombreux* au lieu d'une traduction plus littérale : *de plus en plus de recours*).

Exemple 10 (Europarl)

No decurso do aprofundamento da integração, gerou-se uma lacuna cada vez maior em matéria da protecção dos direitos fundamentais.
L'intégration avançant, une lacune toujours plus grande s'est révélée en matière de protection des droits fondamentaux.(LS : allemand)
O ritmo de aquecimento torna-se cada vez mais rápido.
Ce réchauffement ne cesse de s' accélérer. (LS : néerlandais)

La proportion entre un ensemble et un sous-ensemble se traduit parfois par une fraction, ou un pourcentage :

Exemple 11 (Europarl)
... uma em cada 3 mulheres na República da Moldávia está desempregada.
... un tiers des femmes en Moldavie sont sans emploi. (LS : estonien)
... 20 % da população mundial, dos quais um em cada dois africanos, não tem acesso a água potável.
… 20 % de la population mondiale, et 50 % des Africains, n'ont pas accès à l'eau douce. (LS : espagnol)

La préposition *par* s'utilise aussi parfois pour traduire les deux types de proportion, à condition que deux N interviennent et que Q_2 s'exprime par un N singulier (ici : million), comme dans l'exemple 12.

Exemple 12 (Europarl)
Isto significa que, todos os anos, realizamos 60 transplantes de rim e 30 transplantes de fígado por cada milhão de habitantes.
Ainsi donc, chaque année, 60 reins et 30 foies sont transplantés par million d'habitants.
(Izaskun Bilbao Barandica)

Enfin, lorsque la proportion exprime la fréquence ou la périodicité et que N_2 est une unité de temps, la traduction en français se fait plutôt avec le quantificateur universel *tous les*, quelle que soit la PREP utilisée en portugais :

Exemple 13 (Europarl)
... a cada três segundos, uma criança morre em virtude de uma doença que um médico poderia facilmente prevenir.
… un enfant meurt toutes les trois secondes des suites d'une maladie qu'un médecin pourrait pourtant facilement prévenir. (LS : français)

Ces derniers exemples permettent de justifier ce tableau :

Sens de l'UT	**UCP PT**	**Équivalent FR le plus fréquent**	**Autres solutions acceptables**	**Cas particuliers**
Progression	*cada vez* COMP	*de* COMP *en* COMP	*toujours* COMP	Lexicalisation
Proportion entre un	NUM PREP *cada* NUM	NUM N *sur* NUM		Fraction Pourcentage

ensemble et un sous-ensemble	N (1 N) SQ_1 PREP *cada* SQ_2 (2 N)	SQ_1 *pour* SQ_2	SQ_1 *par* SQ_2	
Proportion entre deux ensembles	SQ_1 PREP *cada* SQ_2	SQ_1 *pour* SQ_2	SQ_1 *par* SQ_2	N_2 = unité de temps → SQ_1 *tou(te)s les* SQ_2

Tableau 1. Modèle de biotraduction.

2.2 Deux systèmes de traduction automatique neuronale

Les deux systèmes de TA neuronale que nous avons testés sont sans doute les plus utilisés aujourd'hui et se passent de longue présentation.

Le premier, Google Translate (GT), a été lancé en 2006, sous le règne de la TA stastistique, afin de « mettre l'information mondiale à la disposition de tous » (Segev 2010). Ce système de TA est passé aux réseaux neuronaux et à l'apprentissage profond en 2016. En 2017, un concurrent de taille est apparu. Il s'agit de DeepL Traducteur (DL) entraîné sur les volumineux corpus alignés de Linguee, dont il est en quelque sorte le prolongement.

Deux ingénieurs du groupe de recherche *Google research* expliquent sur le blogue de *Google AI* les transformations opérées sur GT depuis 2016 pour améliorer les performances du système de TA neuronale, notamment dans les langues disposant de peu de ressources linguistiques (Isaac Caswell & Bowen Liang, 8 juin 2020). Ils mettent toutefois en garde contre les problèmes non résolus, surtout dans les langues en manque de ressources : « poor performance on particular genres of subject matter ("domains"), conflating different dialects of a language, producing overly literal translations, and poor performance on informal and spoken language ». En effet, GT puise la plupart de ses données sur Internet et il n'est pas étonnant que ses données ne couvrent pas les domaines de spécialité, que les variantes linguistiques régionales, telles que le portugais d'Europe et du Brésil, soient mélangées, que la traduction soit quelque peu littérale et que ses performances diminuent lorsqu'il s'agit de traduire un registre informel ou un discours oral.

De son côté, DL dispose d'un volume important de données de qualité et prétend sur son site « concurrencer les entreprises de pointe spécialisées dans l'apprentissage automatique » et s'affirmer « comme une référence grâce à ses percées dans le domaine des mathématiques de réseaux neuronaux et de leur méthodologie ». Les nouveaux modèles lancés en 2020 et en 2021 seraient « capables de restituer encore plus précisément le sens des phrases traitées, et même de traduire haut la main des textes appartenant à divers domaines de spécialité ».

GT semble miser sur le nombre de paires de langues (108 en 2020) tandis que DeepL se limite pour l'instant à 21 langues officielles de l'UE, au russe, au japonais et au chinois, bref à des langues pour lesquelles des corpus alignés de qualité existent.

2.3 Échantillons

Trois échantillons ont été constitués : un échantillon de test et deux échantillons de TA brute. Le premier se compose de 102 occurrences de l'UCP portugaise exprimant la progression, de 42 occurrences de l'UCP exprimant la proportion entre un ensemble et un sous-ensemble et de 24 occurrences de l'UCP exprimant la proportion entre deux ensembles. Cet échantillon n'est pas représentatif de l'usage au Portugal à la fin du XX^{e}

siècle, mais bien de divers emplois comportant divers défis pour la machine. Toutes les occurrences proviennent du corpus CETEMPúblico (CTP). Ce corpus journalistique est annoté, ce qui facilite la précision des requêtes, et compte environ 180 millions de mots écrits pour le journal portugais Público entre 1991 et 1998. C'est un corpus de langue générale qui commence à vieillir. Il reste néanmoins représentatif de l'usage en ce qui concerne ces UCP à caractère idiomatique qui évoluent lentement. Le deuxième et le troisième échantillons contiennent la TA brute par GT et DL de l'échantillon de test qui leur a été soumis en août 2019 puis en septembre 2021.

2.4 Critères d'analyse

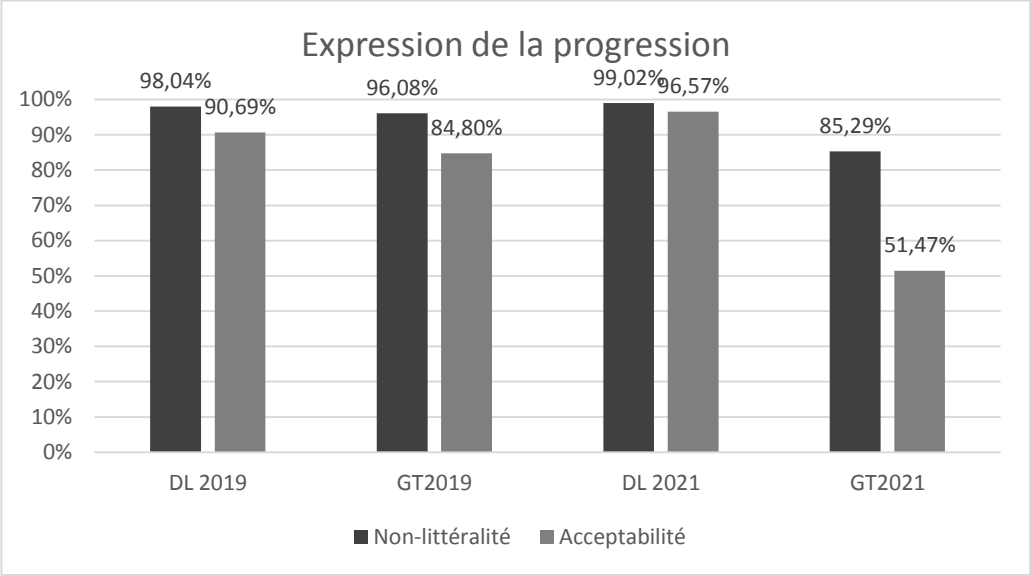
Pour analyser la TA brute, nous avons retenu deux critères généraux. D'une part, la non-littéralté, c'est-à-dire la capacité à identifier les UT et à les traduire en bloc, soit sans *chaque*, et, d'autre part, l'acceptabilité de la proposition de traduction. Pour être acceptable, la TA brute des UT doit faire partie des solutions du modèle de biotraduction, elle doit être sémantiquement juste et syntaxiquement correcte.
Accessoirement, nous avons analysé la capacité de chaque système à relever des défis particuliers tels que la scission et l'inversion et à proposer des solutions originales telles que la lexicalisation.

III ANALYSE DES RÉSULTATS

Entre 2019 et 2021, on pourrait s'attendre à une amélioration sensible des performances de la TA neuronale grâce aux progrès scientifiques et aux efforts des concepteurs concurrents. Alors qu'en est-il de la TA neuronale des trois UT dont il est question ?
Nous présentons et analysons les résultats par UT, en commençant par les critères généraux et quelques exemples justifiant les résultats obtenus par chaque système de TA neuronale. Ces exemples de succès ou d'échec permettent d'analyser la capacité des systèmes à relever des défis particuliers ou à trouver des solutions originales.

3.1 Progression

En ce qui concerne la TA de l'UCP portugaise exprimant la progression, les résultats sont mitigés, comme on peut le voir sur le graphique 1 :



Graphique 1: Résultats de la TA brute de la progression selon les critères généraux.

Les résultats de 2019 sont déjà excellents. GT et DL sont au coude à coude en termes de non-littéralité avec seulement 2% d'écart. L'écart se creuse (environ 6%) en ce qui concerne l'acceptabilité, mais on remarque une baisse de 8% entre le premier et le deuxième critères pour chaque concurrent.
En 2021, DL a encore légèrement progressé et l'écart entre les deux critères n'est plus que de 2,5%. Quant à GT, il dégringole de plus de 10% pour la non-littéralté et d'environ 45% pour l'acceptabilité. Que s'est-il passé ?
Commençons par l'inacceptabilité des solutions proposées par GT en 2021 avec une phrase averbale comportant quatre occurrences coordonnées de *cada vez menos* N, ce qui constitue un fameux défi :

Exemple 14
Cada vez menos votantes, para cada vez menos partidos, para cada vez menos alternativas de escolha, para cada vez menos oportunidades de mudança. (CTP)
De moins en moins d'électeurs, de moins en moins de partis, de moins en moins de choix, de moins en moins de possibilités de changement. (DL 2019)
Moins d'électeurs, moins de partis, moins d'options de choix, moins d'opportunités de changement. (GT 2019)
De moins en moins d'électeurs, de moins en moins de partis, de moins en moins d'alternatives parmi lesquelles choisir, de moins en moins de possibilités de changement. (DL 2021)
* *Moins et moins d'électeurs : pour moins et moins les partis, pour moins et moins alternatives à choisir, pour moins et moins de possibilités de changement.* (GT 2021)

Aucune TA n'est littérale, mais la solution proposée par GT en 2021 est inacceptable. Il s'agit d'un anglicisme qui ne figure pas dans le modèle de traduction contrairement au COMP simple de GT 2019. En 2019 comme en 2021, DL traduit par la solution la plus courante. Ainsi, les quatre systèmes obtiennent le point selon le critère de non-littéralité et GT 2021 est inacceptable. Notons toutefois que la préposition *para* n'a été traduite que par GT 2021, mais l'évaluation se limite à *cada vez* COMP et l'élément éventuellement modifié par COMP. Parfois, GT 2021 se distingue en bien comme dans l'exemple suivant, où le défi résulte de la coordination d'un N et d'un ADJ modifiés par COMP, sans répétition de l'élément ordonnant *cada vez* :

Exemple 15
« *O ambiente e as questões relacionadas com os recursos vão dar origem a cada vez mais conflitos e mais violentos », disse Maurice Strong, …* (CTP)
« L'environnement et les enjeux liés aux ressources vont conduire à des conflits toujours plus nombreux et plus violents », a déclaré Maurice Strong, … (GT 2021)
?*de plus en plus de conflits et de violence* (DL 2019 et 2021)
??des conflits de plus en plus violents (GT 2019)

En recourant à la solution moins fréquente *toujours* COMP, GT 2021 propose la meilleure solution, même s'il s'agit sans doute d'un calque de *ever* COMP, une des solutions possibles en anglais. Tous les autres recourent à la solution la plus fréquente et produisent un sens légèrement décalé (DL) ou omettent l'idée de multiplication des conflits (GT 2019). Notons

que DL conserve la coordination, contrairement à GT 2019, dont la proposition est finalement moins bonne.
L'exemple 16 montre la superiorité de DL 2021, qui parvient à relever un double défi : d'une part, l'étendue de l'UT et, d'autre part, l'humour.

Exemple 16
Há quem, com um certo humor, defina como especialista aquele que sabe cada vez mais de cada vez menos. (CTP)
* *Il y a ceux qui, avec un certain humour, définissent comme un expert celui qui en sait de plus en plus de moins en moins.* (DL 2019
* *Il y a ceux qui, avec un certain humour, définissent comme experts ceux qui en savent de moins en moins.* (GT 2019)
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. (DL 2021)
* *Certaines personnes avec une certaine humeur, définies comme un expert qui en sait toujours plus sur un temps moins.* (GT 2021)

Les quatre systèmes ajoutent le pronom *en* indispensable à cette construction du V savoir, mais seul DL 2021 propose une solution acceptable que l'on peut même considérer comme originale en raison de sa fluidité. DL 2019 et GT 2021 proposent un non-sens et GT 2019 un contresens, trois solutions inacceptables.
La coordination d'UT peut constituer un défi, on l'a vu, c'est aussi le cas de la scission identifiée ci-dessus comme défi syntaxique. L'exemple 17 comporte deux occurrences coordonnées de *cada vez* COMP, dont une est scindée. L'élément ordonnant est séparé de COMP par trois mots, ce qui complique encore la tâche d'identification de l'UT.

Exemple 17
Cada vez as regras imperam menos, cada vez mais tudo é permitido. (CTP)
* *Chaque fois que les règles régissent moins, chaque fois que tout est permis.* (DL 2019)
* *Chaque fois que les règles régissent moins, de plus en plus tout est permis.* (GT 2019)
* *De plus en plus, les règles sont de moins en moins en vigueur, et de plus en plus tout est permis.* (DL 2021)
* *Chaque fois que les règles l'emportent de moins en moins, de plus en plus tout est permis.* (GT 2021)

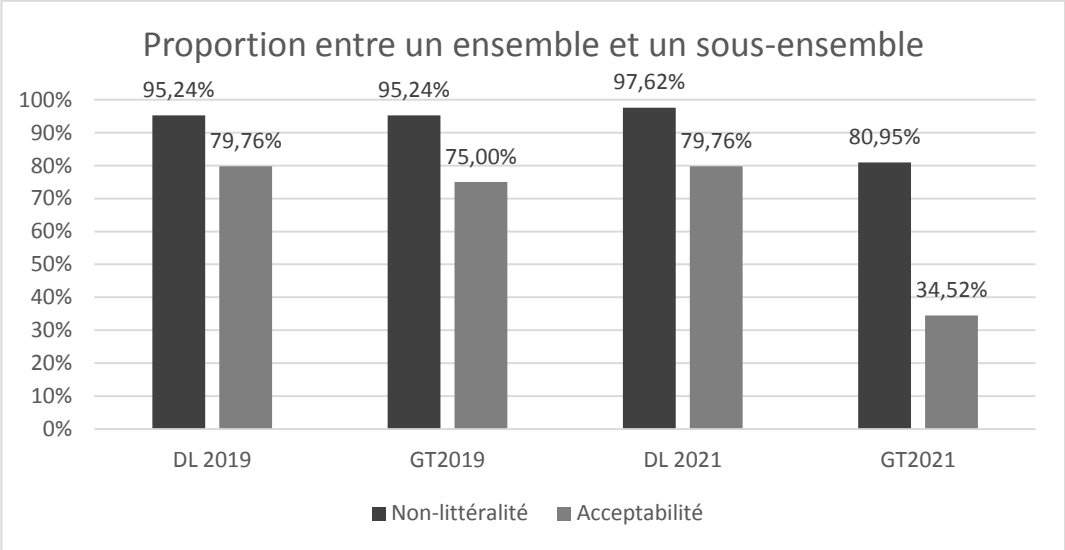
DL 2019 traduit les deux occurrences littéralement en ajoutant le pronom relatif *que* (Theissen 2009) pour former l'UCP *chaque fois que*, ce qui est donc inacceptable. C'est aussi ce que proposent GT 2019 et 2021 pour la première UT, mais ils proposent une solution non-littérale et acceptable pour la deuxième, comme DL 2021. Celui-ci ne propose aucune traduction littérale, mais la première proposition donne lieu à un non-sens.
Pour terminer cette analyse des performances de la TA neuronale de la progression, voyons deux solutions originales proposées par DL :

Exemple 18
Uma obra indispensável, cuja importância se torna cada vez maior. (CTP)
Une œuvre indispensable, dont l'importance ne cesse de croître. (DL 2019 et 2021)
Quebraram todas as regras e estão cada vez melhor. (CTP)

Ils ont enfreint toutes les règles et ils vont de mieux en mieux. (DL 2019)

DL 2019 et 2021 lexicalisent la progression qualitative *se torna cada vez melhor* en proposant *ne cesse de croître* alors que GT propose des solutions acceptables mais moins naturelles en 2019 (*devient de plus en plus grande*) et en 2021 (*croît toujours plus*), où on remarque de nouveau l'influence de l'anglais.
La deuxième solution originale est apportée par DL 2019. Il s'agit d'un expression idiomatique en portugais (*estar bem*/*bom*/*boa*) qui se traduit par *aller bien* plutôt que par les solutions proposées par les autres : ?*sont de mieux en mieux* (DL 2021) ou *s'améliorent* (GT 2019), qui pourrait convenir dans un autre contexte. Quant à GT 2021, sa solution donne lieu à un non-sens : **sont obtiennent mieux*.
En ce qui concerne la progression, l'UCP la plus fréquente en portugais et donc sans doute la mieux représentée dans les données de ces systèmes, le verdict est sans appel : DL s'améliore et GT semble désormais passer par l'anglais pour traduire d'une langue romane à une autre.

3.2 Proportion entre un ensemble et un sous-ensemble
La proportion entre un ensemble et un sous-ensemble est plus rare dans la langue générale. Cela se répercute apparemment sur les résultats, qui sont légèrement plus faibles selon le critère de non-littéralité et nettement en-deçà au niveau de l'acceptabilité. La chute de GT 2021 se confirme. DL progresse légèrement sur la non-littéarilité, mais ses résultats sont stationnaires quant à l'acceptabilité. Cela s'explique sans doute par le manque de données et les nombreuses variables entravant l'identification de l'UT.



Graphique 2 : Résultats de la TA brute de la proportion entre un ensemble et un sous-ensemble

Le premier exemple de TA neuronale brute de l'expression de la proportion entre un ensemble et un sous-ensemble concerne l'ambiguïté résultant de l'emploi de la préposition *em* alors que chaque SQ comporte un N

Exemple 19
… (a prevalência é de um transexual homem em cada 30 mil pessoas, um transexual mulher em cada 100 mil pessoas). (CTP)

… (la prévalence est d'un homme transsexuel pour 30 000 personnes, une femme transsexuelle pour 100 000 personnes). (DL 2019)
* *… (la prévalence est d'un homme transsexuel sur 30 000 personnes, une femme transsexuelle sur 100 000 personnes).* (GT 2019 et 2021)
* *… (la prévalence est d'un transsexuel masculin sur 30 000 personnes, d'une transsexuelle féminine sur 100 000 personnes).* (DL 2021)

La prévalence est une proportion entre un ensemble et un sous-ensemble puisque l'homme et la femme font partie des 30 000 ou 100 000 personnes. D'ailleurs, en portugais, la PREP *em* s'emploie typiquement pour exprimer ce type de proportion. En français, la PREP *sur* s'emploie typiquement avec un seul N et ne convient pas ici. Seule la solution de DL 2019 est acceptable. Notons toutefois que DL 2021 a répété la PREP *de* (*d'une femme transexuelle*), conformément à la norme en français, alors que les autres l'ont tous omise. La prévalence est un terme médical. Il appartient donc à un domaine de spécialité que ces systèmes destinés au grand public ne couvrent pas.
Le défi à relever dans l'exemple 20, c'est l'étendue de N (*clientes das prostitutas estudadas*). Cette caractéristique ne pose pas de problème en portugais, où N se situe dans SQ_2. Par contre, sa place typique en français, c'est dans SQ1, qui s'accomode mal d'un N accompagné de son complément.

Exemple 20
E estima-se, […] que um em cada cinco clientes das prostitutas estudadas seja seropositivo. (CTP)
? […] on estime qu'un client sur cinq des prostituées étudiées est séropositif. (DL 2019 et GT 2019)
? *Et l'on estime […] qu'un client sur cinq des prostituées étudiées est séropositif.* (DL 2021)
* *Et on estime […] que l'un dans cinq clients des prostituées étudiées est séropositive.* (GT2021)

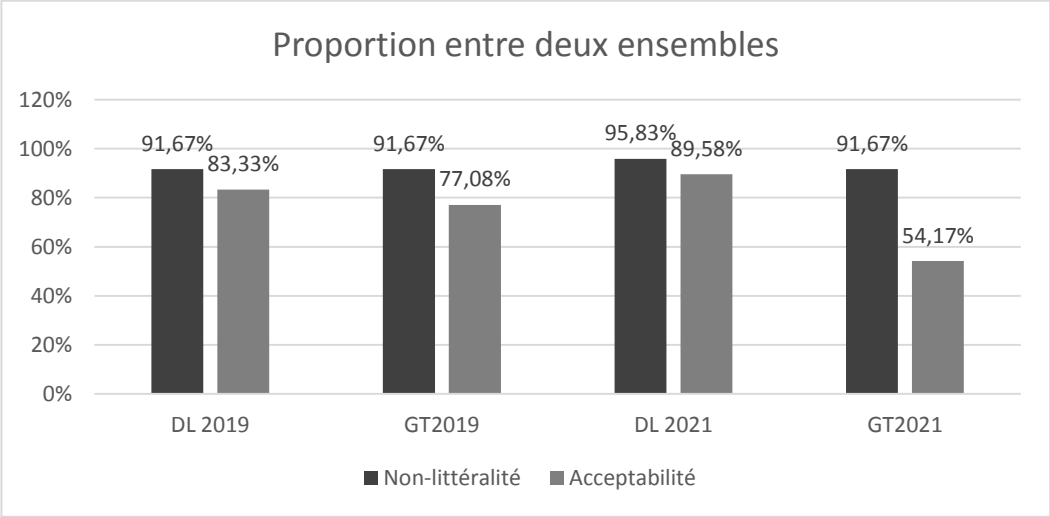
GT 2019, DL 2019 et 2021 séparent N de son complément pour donner lieu à une phrase étrange. Ici, une fraction ou un pourcentage aurait pu rendre le style plus fluide. Quant à GT 2021, son passage par l'anglais se confirme, puisqu'il propose un calque de *one in every five*. En outre, il accorde l'ADJ au féminin singulier, sans doute à cause de la proximité du N féminin *prostituées*.
Dans l'exemple 21, l'inversion se combine à la scission, deux défis syntaxiques difficiles à relever.

Exemple 21
É a história do formigueiro: em cada dez formigas há duas que tematizam o formigueiro, as outras oito não fazem nada... (CTP)
? *C'est l'histoire de la fourmilière : sur dix fourmilières, il y en a deux qui ont pour thème la fourmilière, les huit autres ne font rien ?* (DL 2019)
Voici l'histoire de la fourmilière : sur dix fourmis, deux thématisent la fourmilière, les huit autres ne font rien ... (GT 2019)
C'est l'histoire de la fourmilière : sur dix fourmis, deux s'occupent de la fourmilière, les huit autres ne font rien... (DL 2021)

** Il est l'histoire de la fourmilière : hors de toutes les fourmis dix, il y a deux qui thématiser la fourmilière, les huit autres ne font rien ...* (GT 2021)

GT 2021 ne mérite pas de point. Il n'a pas utilisé *chaque* mais un autre quantificateur universel (*toutes les*) dans son calque de l'anglais *out of every*. La phrase n'a aucun sens. Les autres ont obtenu le point de non-littéralité. Ils ont reproduit l'inversion en faisant passer N dans SQ_2, ce qui est rare en français, mais pas impossible. GT 2019 et DL 2020 ont également obtenu un point pour l'acceptabilité et DL 2019 un demi-point pour avoir confondu *fourmi* et *fourmilière*.
La tendance à passer par l'anglais de GT 2021 se confirme et explique largement ses résultats médiocres. Les progrès de DL ne sont pas aussi spectaculaires que pour la TA de la progression, sans doute parce que la proportion est moins fréquente, qu'elle s'emploie dans des domaines de spécialité non couverts par les données de ces systèmes et que les variables sont plus nombreuses.

3.3 Proportion entre deux ensembles
Dans l'échantillon de test, les occurrences de l'UT exprimant la proportion entre deux ensembles au moyen du quantificateur universel *cada* suivi d'un numéral cardinal sont les moins nombreuses (24). Dans l'ensemble, les résultats de la TA brute sont plutôt meilleurs que ceux de l'autre type de proportion au niveau de l'acceptabilité. DL 2021 a légèrement progressé de quelques pour cent sur les deux fronts et GT 2021 a maintenu son pourcentage de non-littéralité, mais a perdu plus de 20% en acceptabilité.



Graphique 3 : Résultats de la TA brute de la proportion entre deux ensembles.

La supériorité de DL 2021 se manifeste dans l'exemple 21, où l'étendue de SQ_2 (*três metros de área livre*) semble bien constituer un défi.

Exemple 22
… a lotação que, tratando-se de um bar ou discoteca, não deve receber mais do que quatro pessoas por cada três metros quadrados de área livre. (CTP)
* *… quatre personnes pour chaque trois mètres carrés d'espace libre.* (DL 2019)
* *… quatre personnes pour chaque personne. Trois mètres carrés de surface libre.* (GT 2019)
… la capacité, qui, dans le cas d'un bar ou d'une discothèque, ne devrait pas accueillir plus de quatre personnes pour trois mètres carrés de surface libre. (DL 2021)

* *… quatre personnes pour chaque trois mètres carrés de zone libre*. (GT 2021)

Seul DL 2021 propose une solution acceptable. Tous les autres suggèrent une traduction littérale avec *chaque*.
L'exemple 23 illustre la proportion entre deux ensembles avec la PREP *em* au lieu de *por*, plus typique pour ce type de proportion. Aucun système de TA n'a réussi à proposer une solution acceptable.

Exemple 23
A esclerose […] afecta três mulheres em cada dois homens, ... (CTP)
* *La sclérose (en plaques) [...] touche/affecte trois femmes sur deux hommes, …* (DL 2019, GT 2019 et DL 2021)
* *Sclérose en plaques [...] touche trois femmes sur des deux hommes, …* (GT 2021)

Le dernier exemple est court mais les défis sont nombreux : il s'agit ici de fréquence, N_2 représente une unité de temps assortie d'une proposition relative, la structure est inversée et la proportion s'établit entre un événement (*morre uma pessoa vítima do tabaco*) et une unité de temps (*dez segundos*).

Exemple 24
… em cada dez segundos que passam morre mais uma pessoa vítima do tabaco. (CTP)
… toutes les dix secondes, une personne meurt du tabac. (DL 2019)
?… toutes les dix secondes, une personne de plus victime du tabac meurt. (GT 2019)
… une personne supplémentaire meurt du tabac toutes les dix secondes. (DL 2021)
**…dans chaque seconde dix qui passent une autre personne meurt du tabac*. (GT 2021)

Seul GT 2021 suggère une traduction littérale avec *chaque*. La solution de GT 2019 manque cruellement de fluidité. La plus naturelle semble bien être celle de DL 2019. En 2021, DL a reconstitué l'ordre canonique, mais ce n'était pas nécessaire, et il a ajouté *supplémentaire*, moins bon que la *personne de plus* de GT 2019 et un peu redondant.

CONCLUSION
Au terme de cette analyse, nous pouvons conclure que la TA neuronale ne peut toujours pas se passer du biotraducteur, même si celui-ci se transforme souent en plus post-éditeur. Certes, notre échantillon de test ne reflète pas l'usage courant, il contient beaucoup de défis particuliers et peu fréquents, or, les systèmes de TA tendent à généraliser et ont du mal à gérer ce qui sort de l'ordinaire. DL s'améliore, mais ses progrès, spectaculaires lorsque l'on est passé de la TA statistique à la TA neuronale, ralentissent et certains obstacles d'ordre lexical, syntaxique et sémantique persistent, surtout lorsque les UT sont longues, scindées ou lorsque les éléments d'UT binaires sont inversés. La couverture des domaines constitue un autre problème pour les systèmes de traduction 'généralistes'. Ce qui est plus étonnant, c'est le déclin de GT. Comment en est-on arrivé à faire passer la traduction entre deux langues romanes par l'anglais ? C'est sans doute ce qui se passe pour les langues manquant de ressources pour entraîner les systèmes, mais ce n'est ni le cas du français ni du portugais, comme le prouvent les résultats de DeepL. Les systèmes grand public se voient donc confrontés au dilemme entre la qualité, choisie par DeepL et la quantité de paires de langues, choisie par GT.

La TA neuronale fait désormais partie des outils à la disposition du biotraducteur, qui aurait tort de s'en passer. C'est pourquoi la post-édition doit faire partie de la formation des futurs prestataires de services linguistiques pour qu'ils sachent utiliser cet outil formidable à bon escient et en connaissance de cause.

Références bibliographiques

Bacquelaine, F. *Traduction humaine et traduction automatique du quantificateur universel portugais 'cada' en français et en anglais. Étude de phraséologie comparée*. [Thèse de doctorat]. Faculdade de Letras da Universidade do Porto (Porto) 2020.

Carmo, F.: *Post-Editing: a Theoretical and Practical Challenge for Translation Studies and Machine Learning* [Thèse de doctorat]. Faculdade de Letras da Universidade do Porto (Porto) 2017.

Caswell, I. et Liang, B. : Recent Advances in Google Translate [message posté sur le blogue *AI Google* le 8 juin 2020]. https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html [consulté le 19 février 2022]

Cho, K., Van Merriënboer, B., Bahdanau, D. et Bengio, Y.: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *Proceedings of SSST-8*, *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics 2014 ; 103-111. www.aclweb.org/anthology/W14-4012 [Consulté le 14 août 2019].

DeepL : Pourquoi DeepL ? Le traducteur automatique le plus précis et le plus subtil au monde. https://www.deepl.com/fr/whydeepl

Durieux, C. : L'unité de traduction : une unité de sens. In S. Mejri et 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*. Éditions des archives contemporaines (Paris) 2014 ; 381-388.

Granger, S. et Paquot, M. : Disentangling the phraseological web. In S. Granger et F. Meunier (éd.), *Phraseology: An Interdisciplinary Perspective*. Benjamins (Amsterdam / Philadelphia) 2008 ; 27-49.

Kleiber, G. *Tous les, chaque* et *tout* : comment les analyser ?. In L. de Saussure et A. Rihs, (éd.), *Etudes de sémantique et pragmatique françaises*. 2012; VI : 217-259.

Koehn, P. *Neural Machine Translation*. Cambridge University Press (Cambridge), 2020.

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

Leal, A.: Some observations about the quantifier CADA. In M. Villayandre Llamazares (éd.), *Actas del XXXV Simposio de la Sociedad Española de Lingüistica.* Universidad de León 2006 ; 1576-1593.

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

Melby, A. et Warner, T. *The Possibility of Language. A discussion of the nature of language, with implications for human and machine translation*. John Benjamins (Amsterdam), 1995.

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.* 2016 ; Vol.2 : 1635-1649.

Nothomb, A. : *Stupeur et tremblements*. 1999.*Temor e Tremor* (Traduction de Carlos Sousa de Almeida). 2000.
Rocha, P. et Santos, D. : CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*. 2000 : 131-140
Schmale, G. Qu'est-ce qui est préfabriqué dans la langue ? – Réflexions au sujet d'une définition élargie de la préformation langagière. *Langages* 2013 ;189 : 27-45.
Segev, E.: *Google and the Digital Divide. The Bias of Online Knowledge*. 2010.
Theissen, A. Chaque fois/toutes les fois + relative : une construction anticipante particulière. *Revue Romane* 2009 ; 44(2) : 175-194.
Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (éd.), *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012).* http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf [Consulté le 21 février 2022]
Wilks, Y. *Machine Translation. Its Scope and Limits*. Springer (New York) 2009.