Centro de Linguística da Universidade do Porto

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

DeepL et Google Translate face à l'ambiguïté phraséologique

Colloque interdisciplinaire « Vers une robotique du traduire ? »

Université de Strasbourg (UR 1339 LiLPa), 30 septembre et 1er octobre 2021

Françoise Bacquelaine

Universidade do Porto, Faculdade de Letras et CLUP (UIDB/00022/2020 de la FCT)

La TA 2009-2021: qu'en disent les experts?

- « Machine translation is **not,** as some believe, **solved**, **nor** is it **impossible**, as others still claim. » (Wilks 2009 : v)
- « 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)
- Ambiguïté phraséologique = un des obstacles persistants (Koehn 2020)

Objet : 2 UCP (Schmale 2013) et 3 UT

- *cada vez* COMP
  - Progression:
    - *Virou as páginas com um frenesim* ***cada vez maior****.*
    - *Il tourna les pages avec* ***de plus en plus*** *de frénésie.* (AN)
- SQ_1 PREP *cada* SQ_2
  - Proportion entre un ensemble et un sous-ensemble:
    - *Imaginemos que* ***um em cada dois homens*** *fosse objecto de assédio sexual.*
    - *Imaginez qu'****un homme sur deux*** *soit exposé au harcèlement sexuel.* (EP, LS : suédois)
  - Proportion entre deux ensembles:
    - *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****.* (EP, LS: néerlandais)

Défis particuliers de ces UT

- Traduction en bloc d'UCP comportant une ou plusieurs variables comportant un ou plusieurs mots
- Ambiguïté phraséologique
  - *cada vez* COMP: figement moyen (1 seule variable) mais fréquence exceptionnelle du bigramme *cada vez* en PE (Q *de cada vez* (Q *à la fois*), *de cada vez* (*chaque fois* ou *à chaque fois*), (*de*) *cada vez que* (*chaque fois que* ou *à chaque fois que*)
  - SQ_1 PREP *cada* SQ_2: figement très faible (nombreuses variables), deux UT de même structure, mêmes PREP possibles en PT, beaucoup moins fréquentes/disséminées que la progression

Défis syntaxiques

- Scission (progression)
  - ***Cada vez*** *se torna* ***mais indispensável e insubstituível*** *uma política de verdade, ...* (CTP)
- Inversion des SQ avec ou sans scission (proportion entre un ensemble et un sous-ensemble)
  - *… no tempo dos portugueses* ***em cada 100 casas dez*** *eram para os timorenses …* (CTP)
  - *Em média,* ***em cada 300 preservativos*** *há* ***um*** *que se rompe, …* (CTP)
- Inversion des SQ avec scission (proportion entre deux ensembles)
  - ***Por cada seis dias seguidos de trabalho*** *tem direito a* ***um dia de descanso****; ...* (Sindeq, 2014)

Méthologie

- Corpus parallèles (AN) et alignés (EP) PE-FR → Modèle de biotraduction
- Corpus PE → Échantillon de test représentatif de la diversité d'emplois potentiels et non de l'usage le plus courant
  - Progression: 102 occurrences
  - Proportion entre un ensemble et un sous-ensemble: 42 occurrences
  - Proportion entre deux ensembles: 24 occurrences
- Outils: Google Translate et DeepL
- Échantillon testé en août 2019 et en septembre 2021
- Analyse de la TA brute d'après le modèle de biotraduction

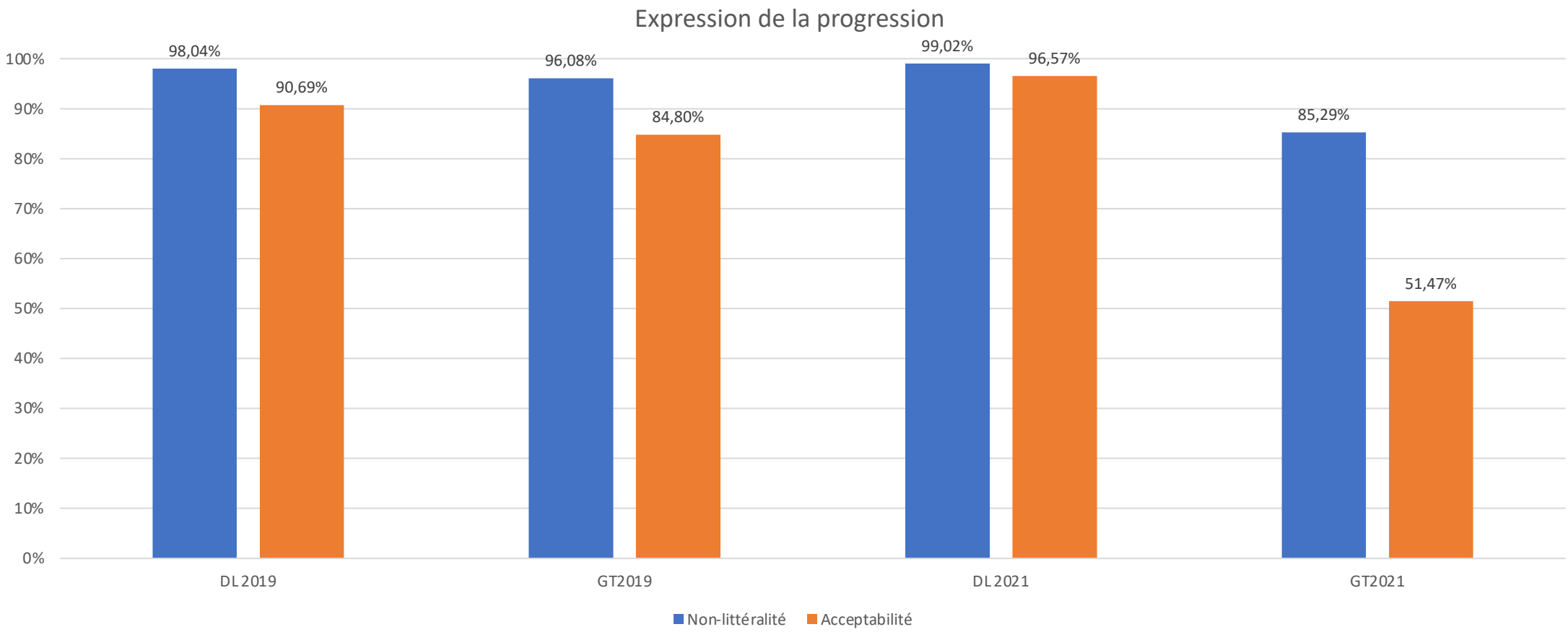
Modèle de traduction

Sens de l'UT	UCP portugaise	Équivalent FR le plus fréquent	Autres solutions acceptables	Cas particuliers
Progression	*cada vez* COMP	*de* COMP *en* COMP	*toujours* COMP (simple COMP)	Lexicalisation
Proportion entre un ensemble et un sous-ensemble	NUM PREP *cada* NUM N (1 N)	NUM N *sur* NUM		Fraction
	SQ_1 PREP *cada* SQ_2 (2 N)	SQ_1 *pour* SQ_2	SQ_1 *par* SQ_2	Pourcentage
Proportion entre deux ensembles	SQ_1 PREP *cada* SQ_2	SQ_1 *pour* SQ_2	SQ_1 *par* SQ_2	N2 = unité de temps → SQ_1 tou(te)s les SQ_2

Critères d’analyse

- Critères généraux
  - Non-littéralité: traduction en bloc sans *chaque*
  - Acceptabilité: une des solutions du modèle + syntaxe + sens
- Capacité à relever des défis particuliers
- Solutions originales

Résultats *cada vez* COMP



Exemple d’inacceptabilité de GT 2021

- *Cada vez menos votantes, para cada vez menos partidos, para cada vez menos alternativas de escolha, para cada vez menos oportunidades de mudança.* (CTP; 4 occurrences; défi: coordination d’UT; phrase averbale)
- *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)

Exemple de succès de GT 2021

- *«O ambiente e as questões relacionadas com os recursos vão dar origem a **cada vez mais conflitos e mais violentos**», disse Maurice Strong, …* (CTP; défi: coordination de N et ADJ, sans répétition de l'él. ordonnant *cada vez*)
- *« 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)

Exemple de supériorité de DL 2021

- *Há quem, com um certo humor, defina como especialista aquele que **sabe cada vez mais de cada vez menos**.* (CTP; défi = étendue de l'UT)
- ** 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) – Non-sens
- ** Il y a ceux qui, avec un certain humour, définissent comme experts ceux qui **en savent de moins en moins**.* (GT 2019) - Contresens
- *Il y a ceux qui, avec un certain humour, définissent le spécialiste comme celui qui **en sait de plus en plus <u>sur</u> de moins en moins <u>de choses</u>**.* (DL 2021)
- ** Certaines personnes avec une certaine humeur, définies comme un expert qui **en sait toujours plus sur un temps moins**.* (GT 2021) – Non-sens

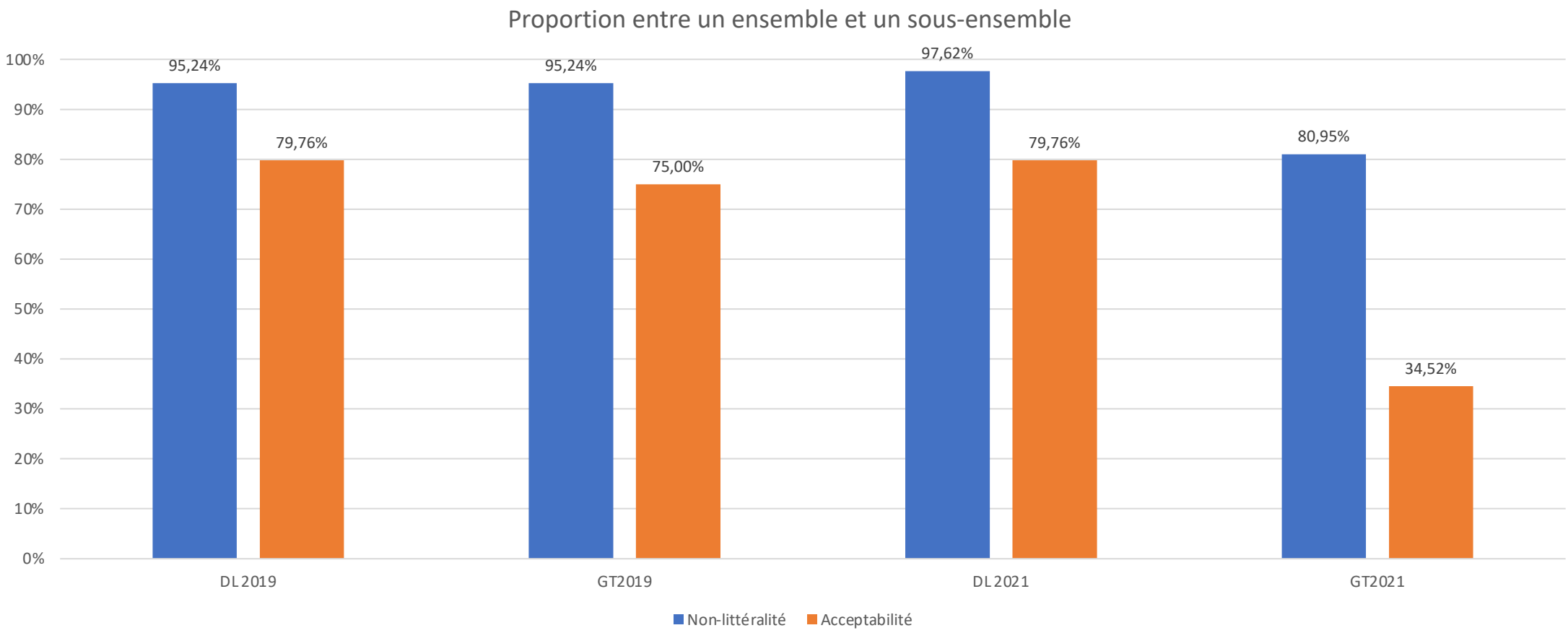
Exemple de scission mal gérée en 2019 et en 2021

- ***Cada vez*** *as regras imperam* ***menos****, cada vez mais tudo é permitido.* (CTP; coordination de deux occurrences dont une scindée)
- * ***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)

Exemples de solutions originales

- *Uma obra indispensável, cuja importância* ***se torna cada vez maior****.* (CTP; défi: transposition de l'ADJ PE en V FR)
- *Une œuvre indispensable, dont l'importance* ***ne cesse de croître****.* (DL 2019 et 2021)
- *Quebraram todas as regras e* ***estão cada vez melhor****.* (CTP; défi: idiomatisme)
- *Ils ont enfreint toutes les règles et ils* ***vont de mieux en mieux****.* (DL 2019)

Résultats: SQ_1 PREP *cada* SQ_2 (ESE)



Ambiguïté résultant de la préposition

- *… (a prevalência é de* ***um transexual homem em cada 30 mil pessoas, um transexual mulher em cada 100 mil pessoas****).* (CTP; défi: PREP *em* avec 2 N)
- *… (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)

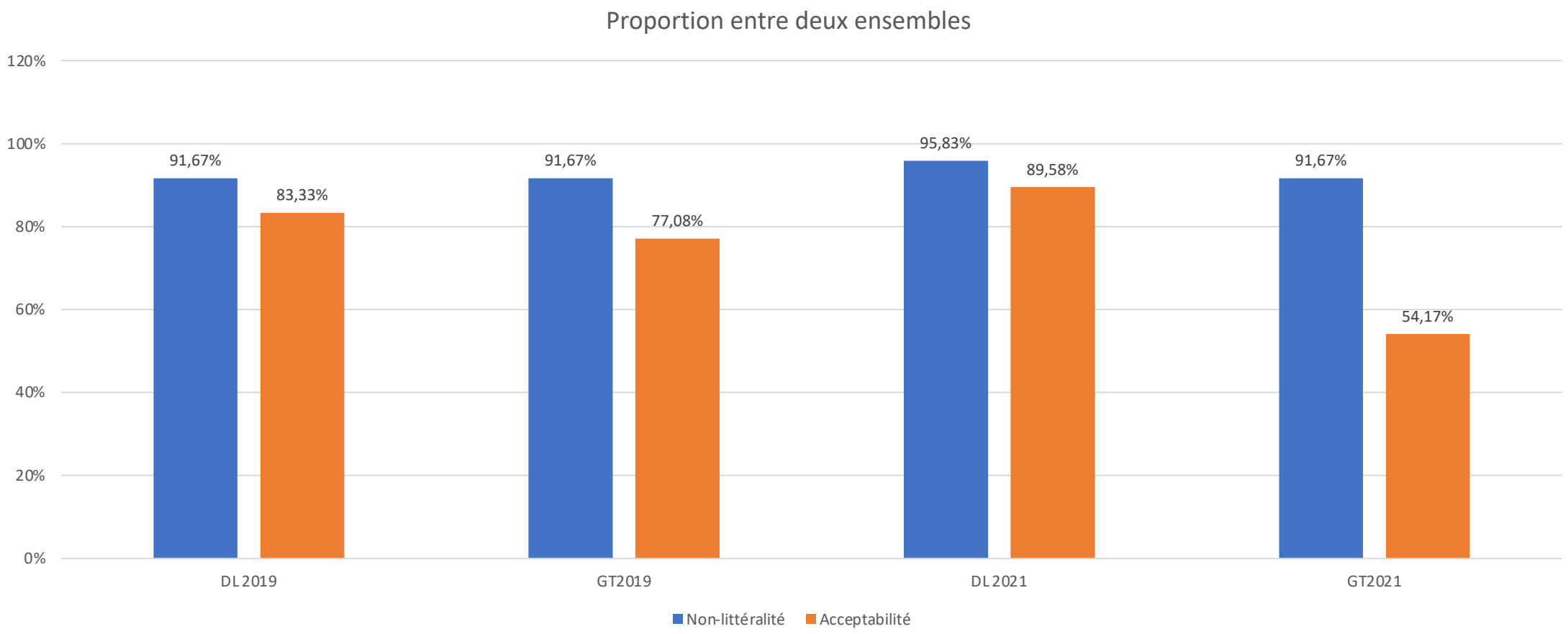
Défi syntaxique: place de N

- *E estima-se,[…]que* ***um em cada cinco clientes das prostitutas estudadas*** *seja seropositivo.* (CTP; 1 N et longueur de SQ_2 en PE)
- ? […] *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)

Inversion et scission

- *É a história do formigueiro:* ***em cada dez formigas*** *há* ***duas que tematizam o formigueiro****, as outras oito não fazem nada...* (CTP; inversion et scission)
- ? *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)

Résultats: proportion entre deux ensembles



Ambiguïté résultant de la préposition

- *A esclerose* […] *afecta* ***três mulheres em cada dois homens,***
- * *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)

Supériorité de DL 2021

- *… 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: défi: étendue de SQ_2)
- * … ***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)

Inversion et scission

- … ***em cada dez segundos que passam morre mais uma pessoa vítima do tabaco.*** (CTP; défis: PREP *em*; étendue des SQ; SQ_2: unité de temps /SQ_1: événement)
- … ***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) – Manque de fluidité
- … ***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)

Conclusion

- En 2021, DL s'est légèrement amélioré ou a maintenu son niveau de 2019.
- GT a nettement régressé par rapport à 2019.
- Divers indices suggèrent que GT utilise désormais l'EN comme langue pivot entre les deux langues romanes que sont le PE et le FR.
- L'échantillon est riche en défis particuliers, or la TA tend à généraliser, GT peut-être plus que DL.
- La TA fait désormais partie des outils à la disposition du biotraducteur et l'enseignement/apprentissage de la post-édition aux futurs traducteurs est devenu indispensable.

Quelques sources

- 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]. Faculdade de Letras da Universidade do Porto.
- Carmo, F. (2017). *Post-Editing: a Theoretical and Practical Challenge for Translation Studies and Machine Learning* [Thèse de doctorat]. Université de Porto.
- Cho, K., Van Merriënboer, B., Bahdanau,D. & Bengio, Y. (2014), On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103-111). Association for Computational Linguistics. <www.aclweb.org/anthology/W14-4012> [Consulté le 14 août 2019].
- Durieux, Christine (2014). L'unité de traduction : une unité de sens. In S. Mejri & 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.
- Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (éd.), *Phraseology: An Interdisciplinary Perspective* (pp. 27-49)
- Kleiber, G. (2012). *Tous les, chaque* et *tout* : comment les analyser ?. In L. de Saussure, & A. Rihs, (éd.), *Etudes de sémantique et pragmatique françaises* (pp. 217-259). Peter Lang

Quelques sources

- Koehn, P. (2021, 17 juin). *Applying New Advances in AI-based Machine Translation to Real World Use* [Webinar]. Omniscien.com . https://omniscien.com/webinars/applying-new-advances-in-ai-based-machine-translation-to-real-world-use-cases/
- Koehn, P. (2020). *Neural Machine Translation*. Cambridge University Press.
- Leal, A.(2006). Some observations about the quantifier CADA. In M. Villayandre Llamazares (éd.), *Actas del XXXV Simposio de la Sociedad Española de Lingüistica* (pp. 1576-1593). Universidad de León.
- 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 (355-366). Honoré Champion.
- 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.
- Scott, B. (2018). *Translation, Brains and the Computer. A Neurolinguistic Solution to Ambiguity and Complexity in Machine Translation*. (Série *Machine Translation: Technologies and Applications* (Vol. 2). Springer International Publishing.
- Wilks, Y. (2009). *Machine Translation. Its Scope and Limits*. Springer