**Variation terminologique et traduction automatique neuronale : étude et évaluation du traitement de la variation terminologique en santé, environnement, traitement automatique des langues**

**Codirection : Pr Natalie Kübler ; Dr Alexandra Mestivier, CLILLAC-ARP, Université Paris Cité**

**Présentation du projet**

**Contexte scientifique**

La traduction neuronale *(neural machine translation)* a émergé à la fin des années 2010 à la suite des progrès fulgurants réalisés dans le domaine de l’apprentissage automatique avec le développement des réseaux de neurones grâce à la technique du *deep learning.* Si les fondements théoriques de cette technologie remontent aux années 1940, c’est l’accroissement des données disponibles, des capacités de stockage et de calcul des processeurs, au début du XXIe siècle qui a permis leur usage massif (Dreyfus 2019 ; Ganascia 2019). Depuis 2016, cette nouvelle technique s’est rapidement imposée en surpassant, lors des évaluations automatiques, basées sur des métriques, les performances des systèmes statistiques et des systèmes hybrides qui étaient la référence jusqu’alors (Wu *et alii.* 2016 ; Bentivogli *et alii.* 2016).

Les systèmes de traduction automatique neuronale (TAN) ont fait de considérables progrès et ne sont plus utilisés désormais que dans le monde de la traduction, mais aussi, dans celui de la rédaction en anglais par exemple (Goulet et al. 2017) ou dans l’enseignement des langues (Bourdais 2021). Si le développement des systèmes de TAN s’est accompagné d’améliorations notables en matière de fluidité, de nombreux points restent encore à améliorer de manière indispensable. Pour Yvon (2019), on assiste à l’apparition d’un « e-traductionnais », nommé aussi « machine translationese » (Bizzoni et al. 2020). Par ailleurs, la plupart des systèmes traduisant segment par segment, la prise en compte du contexte au-delà de la phrase pose encore problème.

Dans les langues de spécialité, les caractéristiques propres à celles-ci, ce qui les distinguent de la langue générale, engendrent d’autres types de problèmes dans la traduction automatique. Le traitement de la terminologie dans les discours spécialisés mérite amplement d’être amélioré.

Bordet et al. (2023) soulignent ainsi les difficultés de traduction des groupes nominaux complexes anglais/français, phénomène emblématique du discours scientifique actuel, dans lesquels, des questions terminologiques peuvent se poser. La traduction des groupes nominaux complexe comporte effectivement différents types de difficultés liées à l’analyse en constituants, la reconnaisses de termes au sein du groupe nominal, l’identification de la tête du groupe, la factorisation des modifieurs ou encore, les noms de marque ou les sigles.

Un autre phénomène, celui de l’incohérence textuelle terminologique et syntaxique peut lui aussi se manifester. On peut ainsi observer une cooccurrence de formes correctes et de formes incorrectes pour différents équivalents d’un terme source en langue cible, ou au contraire des variations morphologiques et structurales de termes complexes dans un document (Daille 2017).

Les récentes recherches sur la variation terminologique ont montré que celle-ci ne fonctionne pas seulement comme un signe de néologie, mais peut être motivée cognitivement (Tercedor 2011) ou rhétoriquement (Gledhill & Pecman 2018). L’une des formes de la variation, l’alternance compacté/décompacté (p. ex., N N et N of N en anglais), se produit différemment selon les parties du texte, en fonction d’objectifs de communication spécifiques. Ainsi, l’alternance entre forme compactée dans le titre et forme décompactée dans l’introduction d’un article, permet aux auteurs de mettre en lumière l’aspect innovant de leur recherche. Une seconde question essentielle, encore peu étudiée, concerne la génération de nouveaux termes en français par création lexicale, s’appuyant sur des processus connus de la formation secondaire entre les deux langues ou encore sur l’émergence de nouveaux modèles de formation (tels que les GNC avec des pré-modifieurs adjectivaux, cf. Mestivier 2015) et qui sont rendues nécessaires par les avancées rapides de la science (Langlais et al. 2009).

**Sujet de thèse et contexte de travail**

Le ou la candidate contribuera à mieux comprendre et mieux décrire la variation terminologique dans des contributions scientifiques dans trois domaines :

* Sciences de la Terre, de l’Environnement et de la Planète
* Traitement Automatique des Langues
* Santé.

La description de la variation se fera non seulement au niveau du texte, mais aussi des différents « moves » du document (Swales Move Analysis), avec pour objectif de permettre un meilleur traitement de la variation au niveau de la traduction automatique. Pour ce faire, le ou la candidate travaillera à partir de corpus monolingues spécialisés dans les trois domaines, en français et en anglais (corpus comparables), mais aussi de corpus alignées anglais/français - corpus de traduction). Le ou la candidate participera aussi à l’annotation des sorties de TA et des post-éditions effectuées par des spécialistes du domaine (Kübler et alXXX ) .

Cette recherche s’inscrit dans la problématique du projet MaTOS (Machine Translation for Open Science), dirigé par François Yvon (ISIR, CNRS Paris Sorbonne) auquel participent le CLILLAC-ARP (Université Paris Cité), l’INIST et l’INRIA.

Le projet MaTOS (<https://anr-matos.github.io/>) a pour objectif de développer de nouvelles méthodes de traduction automatique intégrale de documents et de proposer de nouvelles métriques pour analyser la qualité des traductions produites par le système. Le projet vise à traduire automatiquement des articles scientifique anglais/français dans les trois domaines cités plus haut. MaTOS s’inscrit dans le mouvement de la science ouverte et dans l’effort de diffusion scientifique multilingue qui soutient la traduction automatique post-éditées par des spécialistes, dans différentes langues.

**Tâches**:

* Participation à la validation de candidats-termes extraits des corpus dans les différents domaines en collaboration avec les membres du projets MaTOS ;
* Choix d’un échantillon de textes représentatifs et récents dans les trois domaines ;
* Analyse et description de la variation terminologique dans l’échantillon, afin d’obtenir un modèle ;
* Test du modèle sur d’autres échantillons ;
* Contribution à la recherche sur la variation terminologique dans les articles scientifiques et en traduction automatique
* Rédiger les résultats dans des articles en co-auteurs ou seul-e avec les co-directrices de thèse ;
* Rédiger la thèse.

**Compétences demandées**

* Master en traduction, linguistique anglaise, linguistique appliquée, linguistique de corpus, langues de spécialité, terminologie, ou autres en lien avec le sujet de la thèse.
* Connaissance des outils de linguistique de corpus.
* Excellente maîtrise de l’anglais ou du français et au minimum C1 pour l’autre ; ou langue première en français ou anglais et C1 pour l’autre ; ou Master en études anglophones en France ou en français à l’étranger.
* Très bons résultats en master.
* Capacité à travailler de manière autonome et à prendre des initiatives ; aptitude à travailler en équipe

**Candidater**

**Envoyer à** **natalie.kubler@u-paris.fr** **et** **alexandra.mestivier@u-paris.fr**

* Lettre de motivation expliquant en quoi le sujet et le projet MaTOS vous intéressent
* 1 CV
* Relevé de notes de M1 et M2 (ou un document attestant de votre réussite)
* Mémoire de M2
* Éventuellement lettre de recommandation

Date-limite d’envoi : 6 juillet 2023

**Bibliographie**

* Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo & Marcello Federico. 2016. “Neural versus Phrase-Based Machine Translation Quality: a Case Study”. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing [en ligne]. Austin, Texas: Association for Computational Linguistics. novembre 2016, 257‑267.
* Bizzoni, Yuri Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. In Proceedings of the 17th International Conference on Spoken Language Translation, 280–290, Online. Association for Computational Linguistics.
* Bourdais, Aurélie. 2021 “Outils d’aide à la traduction : pratiques numériques ordinaires en contexte scolaire”. Lidil. Revue de linguistique et de didactique des langues, (63).
* Dreyfus, Gérard. 2019. Réseaux de neurones. [en ligne]. Encyclopædia Universalis. <http://www.universalis-edu.com.rproxy.sc.univ-paris-diderot.fr/encyclopedie/reseaux-de-neurones-formels/>
* Ganascia, Jean-Gabriel. 2019. “Intelligence artificielle. (IA). [en ligne]. Encyclopædia Universalis. <http://www.universalis-edu.com>
* Gledhill, Christopher & Mojca Pecman. 2018. “On alternating pre-modified and post-modified nominals such as aspirin synthesis vs. synthesis of aspirin: Rhetorical and cognitive packing in English science writing”. Fachsprache-Journal of Professional and Scientific Communication, 40(1 + 2), 24-46.
* Goulet, Marie-Josée, Michel Simard, Carla Parra Escartin & Sharon O’Brien. 2017. “La traduction automatique comme outil d’aide à la rédaction scientifique en anglais langue seconde : résultats d’une étude exploratoire sur la qualité linguistique”. ASp. la revue du GERAS. 1 novembre 2017. N° 72, pp. 5‑28. DOI 10.4000/asp.5045.
* Mestivier-Volanschi, A. (2015). Productivity and diachronic evolution of adjectival and participial compound pre-modifiers in English for specific purposes. *Fachsprache*, *37*(1-2), 2-23.
* Swales, John M. *Genre Analysis: English in Academic and Research Settings*. New York: Cambridge University Press, 1990
* Tercedor, M. (2011). The cognitive dynamics of terminological variation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *17*(2), 181-197.
* Yonghu Wu, Mike Schuster, Zhifeng Chen, Quoc Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yua Caon, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gows, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, VINYALS, Oriol Vinyals, Greg Corrado, Macduff Hughes & Jeffrey Dean. 2016. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. arXiv:1609.08144 [cs] [en ligne]. 8 octobre 2016.
* Yvon, François. 2019. “Les deux voies de la traduction automatique”. Hermes, La Revue. 28 novembre 2019. Vol. n° 85, n° 3, 62‑68.