PROPOSITION SUJETS DE THESE
CONTRATS DOCTORAUX
2025—2028

x Contrat doctoral fléché FR Agorantic

x Contrat doctoral fléché EUR InterMEDIUS

Dans le cas où vous souhaitez candidater sur les deux contrats doctoraux, veuillez cocher les deux cases.

Directeur de thèse :	**Pr Fabrice Lefèvre**	**Mail :**	**Fabrice.lefevre@univ-avignon.fr**
Laboratoire :	**LIA**	**Téléphone :**	**0698684501**
Co-directeur :	**Pr Laurent Lombard**		
Laboratoire :	**ICTT**		

Titre en français : L'écart en traduction : Compréhension, gestion et traitement des écarts linguistiques et culturels par l'intelligence artificielle

Titre en anglais : The translation gap: Understanding, managing and processing linguistic and cultural gaps using artificial intelligence

Résumé (7 lignes maximum) : Il est admis que la traduction est un processus complexe où les différences linguistiques, culturelles et contextuelles entre les langues source et cible peuvent engendrer des "écarts". Lexicaux, grammaticaux, syntaxiques ou culturels, ces écarts peuvent perturber la fidélité, la fluidité ou la pertinence du texte traduit. D'où la notion de création et d'art de la traduction.

Dans un contexte de mondialisation et de digitalisation croissante liée à l'intelligence artificielle, notamment via les systèmes de traduction automatique de type LLM, la thèse entend s'intéresser à la question scientifique : l'IA peut-elle être un outil pour la gestion des écarts de traduction ?

Summary: Translation is a complex process in which linguistic, cultural and contextual differences between the source and target languages can give rise to "gaps". Lexical, grammatical, syntactic or cultural, these gaps can disrupt the fidelity, fluidity or relevance of the translated text. Hence the notion of creation and the art of translation.

In a context of globalisation and increasing digitalisation linked to artificial intelligence, in particular via LLM-type machine translation systems, the thesis aims to address the scientific question: can AI be a tool for managing translation gaps?

Mots clés : human translation, stochastic machine translation, translation gaps, large language models

Keywords : human translation, stochastic machine translation, translation gaps, large language models

1- Présentation du sujet (3 pages maximum)

La relation entre traduction et intelligence artificielle est déjà ancienne. Les premiers programmes d'IA des années 50 avaient déjà la traduction entre langues naturelles parmi leurs objectifs prioritaires. C'est d'ailleurs leur échec relatif sur cette question qui mènera au premier hiver de l'IA des années 80 et 90. Avant que les approches probabilistes, ou distributionnelles, viennent relancer fortement le sujet avec des résultats très prometteurs au début du siècle. Résultats qui seront largement amplifiés avec l'avènement des modèles neuronaux de type à « apprentissage profond », que permettent enfin les capacités de calculs des années 2010 (et l'apparition des processeurs graphiques). Sans prétendre faire un historique de l'IA, il est important de rappeler à quel point la traduction automatique est au cœur de son développement. Ce que souligne d'ailleurs de façon imparable le fait que le modèle à l'origine de ChatGPT et consorts, le Transformer, fut initialement développé pour la traduction [Vaswani, 2017].

Depuis cet avènement, la combinaison vertueuse amélioration des performances et grandes disponibilité permet de généraliser son usage pour une multitude de tâches annexes. Par exemple l'utilisation de l'IA pour la « localisation » de systèmes de conversation humain-machine (prise en compte de particularités culturelles, géographiques, sociales...) avec des traductions automatiques [Njifenjou et al, 2023, 2023a, 2025]. La coopération entre l'humain et l'IA n'est pas limitée à la fourniture des données initiales (corpus bi-textuels) par les humains. Elle peut être complétée en interaction (*humain-in-the-loop*), comme par l'approche de post-édition (par exemple [Rubino et al, 2012]) ou encore la compilation de mémoires de traduction pour favoriser l'adaptation à des particularités très marquées (ou simplement mal gérées automatiquement). Ces apports continuent malgré tout d'être discutés (par exemple une comparaison traduction humaine vs traduction auto avec post-édition qui n'est pas en faveur de cette dernière).

Le débat en cours, tant au plan national qu'international, porte donc sur l'intérêt et le niveau réel de la qualité de ces traducteurs automatique. Comment sont mesurées ces performances ? En se comparant généralement à une référence humaine(ou vérité terrain), à l'aide d'une distance approximative (le BLEU, se basant globalement sur un taux de sous-séquences communes, de tailles plus ou moins grandes, de 1 à 5 mots). Alors la comparaison avec la traduction humaine continue de se poser : Human Translation vs. Machine Translation: Which Is Better?, Machine translation VS Human translation: A comparative Analysis, Comparing machine translation and human translation: A case study.

Une solution d'apparence paradoxale est l'utilisation de l'IA comme évaluatrice, « LLM-as-a-judge », de la traduction humaine. Cette utilisation (provocatrice ?) de l'IA offre malgré tout des performances appréciables [Benchmarking GPT-4 against Human Translators: A Comprehensive Evaluation Across Languages, Domains, and Expertise Levels] et surtout d'un niveau tout à fait utilisable si on se place dans une perspective de coopération avec l'IA, plutôt que dans un simple remplacement.

On peut alors s'attacher à regarder plus en détails des phénomènes complexes, y compris pour les traducteurs. Est-elle capable de détecter les situations d'intraduisibles (selon la définition de l'académicienne B. Cassin) ou de difficultés en général, tout autant que d'expliquer les processus de généralisation sous-jacents (puisqu'il n'y pas d'explicitation, sous forme de règles compilées, telle une grammaire) ? Et cela aussi bien aux niveaux lexical, syntaxique, que pragmatique. Des travaux récents s'attachent à fournir des outils pour appréhender ce sujet, comme « Holmes: A Benchmark to Assess the Linguistic Competence of Language Models, » qui inspirent notre réflexion.

Dans la tension qui se joue entre traduction artisanale et traduction automatique, et plus largement entre Intelligence artificielle et Intelligence humaine en tant que double énigme, pour reprendre le titre de l'ouvrage de Daniel Andler (sa critique ici), la question des intraduisibles semble essentielle dans ce travail de recherche. Si l'on se réfère aux travaux de l'académicienne Barbara Cassin, la théorie des

intraduisibles est une approche philosophique et linguistique qui explore la nature des concepts et des mots qui ne peuvent pas être pleinement traduits d'une langue à une autre, en raison de leurs particularités culturelles, historiques ou philosophiques. Dans son ouvrage « La Guerre des traductions » (2004), Cassin introduit l'idée que certains termes, en particulier dans les domaines de la philosophie, du droit, ou de la culture, sont profondément ancrés dans des contextes spécifiques, ce qui rend leur traduction exacte impossible. Ces « intraduisibles » ne se contentent pas d'être des mots difficilement traduisibles, mais désignent des concepts qui portent des significations uniques liées à des traditions et des systèmes de pensée particuliers. La théorie des intraduisibles montre également que chaque langue véhicule des visions du monde particulières. Ces visions ne peuvent pas toujours être capturées par une autre langue, ce qui peut mener à des malentendus ou à des approximations, mais aussi à une richesse interprétative. Cassin insiste sur le fait que la traduction n'est pas un simple transfert de mots d'une langue à une autre, mais un acte interprétatif, où le traducteur en tant que créateur joue un rôle crucial : il doit décider comment rendre le sens tout en étant conscient de l'impossibilité d'une traduction parfaite. Si le traducteur artisanal est dès lors naturellement confronté à une tension entre fidélité et créativité, qu'en est-il de la traduction automatique ?

Pour avancer dans cette double interrogation sur la nature des traductions humaine/artisanale et automatique/artificielle, le travail envisagé se base sur **la notion d'écart**. L'écart est le phénomène principal dans la traduction qui permette au fond de parler de création [Masson, 2017]. C'est parce qu'elle est une œuvre de trahison (l'écart trahit naturellement l'œuvre originale) qu'elle devient une œuvre de création. La nécessité de tenir compte des apports créateurs des traductions, et en particulier des apports créateurs des écarts en traduction, a pour conséquence d'une part d'ouvrir un chantier vaste et inédit ; d'autre part d'interroger la notion de « patrimoine » en tant que les traductions enrichissent les langues et tous les domaines du savoir [Chevrel, 1997].

L'importance de l'écart, qui est aussi confrontation, vient de ce qu'il rend possible à la fois une tension créatrice dans l'acte de traduire et s'oppose au fond à toute idée d'uniformisation et d'homogénéisation : « L'universalité est un concept de la raison, l'uniforme un concept de la production : celle-là invoque une nécessité, celui-ci ne repose que sur une commodité » F. Jullien, Du mal/du négatif, Paris, Seuil, 2006, p. 18.

A partir de là, nous souhaitons réfléchir sur les rapports entre l'IA et la traduction. Après avoir établi un **diagnostic des écarts**, on peut alors se poser la question : l'IA peut/pourra-t-elle résoudre ces écarts et sous quelle(s) forme(s) et condition(s) ? Ainsi, la problématique fondamentale de la thèse pourrait être : « Dans quelle mesure l'IA peut-elle être un outil pour la gestion des écarts de traduction ?»

Nous pouvons distinguer deux phases principales pour élaborer ce questionnement :

A) Phase de diagnostic : les algorithmes actuels, bien que performants dans de nombreux cas, peinent à capturer toute la richesse et la complexité des nuances culturelles et linguistiques en traduction, souvent en raison de leurs limitations dans l'analyse contextuelle, des ambiguïtés sémantiques ou de la prise en compte insuffisante des spécificités du texte source et surtout des émotions suscitées par les mots. Avec une étude précise de la littérature confortée avec des expériences locales, un corpus d'écarts représentatifs sera constitué, pour un nombre suffisant de paires de langues (parmi lesquelles au moins : anglais, français, italien, grec, russe, chinois). A titre d'exemple, une discussion récente lors d'une journée d'études organisée à la Cité internationale de la langue française, des encadrants avec l'auteur-traducteur André Markowicz, célèbre pour avoir traduit toute l'œuvre de Dostoïevski ou le théâtre de Gogol, Tchekhov, nous a convaincu de l'importance de ces écarts porteurs de la créativité de l'auteur-traducteur lorsqu'il décide s'en emparer, et rappelé que cela lui a valu de fortes réactions à la réception de ses premières traduction de Dostoïevski dans les années 1990. Il estime que les traductions originales ont fait fausse route, car « Dostoïevski détestait l'élégance, en particulier celle des Français. Il écrivait avec véhémence, sans se soucier de la syntaxe ni des répétitions ». Le corpus comparatif des traductions de cet auteur est donc une toute première source de données pour aider à révéler la nature des ces écarts.

B) Phase d'élaboration des outils et méthodes : explorer la notion d'écart en traduction permet de mettre en lumière les différents types d'écarts rencontrés, ainsi que les stratégies actuelles de traitement de ces écarts par l'intelligence artificielle et d'en établir la (dys)fonctionnalité dans le texte. Il s'agit alors :

- d'interroger comment l'intelligence artificielle peut identifier, analyser et prendre en charge les écarts en traduction tout en respectant les exigences de précision, de fluidité et de fidélité culturelle du texte traduit ? Au-delà de l'état de l'art des solutions adaptées seront envisagées et étudiées, passant par l'adaptation fine des modèles existants ou le recours à la combinaison (*mixture of*) d'experts. Un mécanisme important envisagé est le recours à la multi-traduction. En effet les écarts s'instancient dans une paire de langues particulière mais se propagent ensuite dans les éventuelles traductions suivantes (il est notable que beaucoup d'œuvres littéraires, par exemple de langues plus minoritaires, sont d'abord traduites dans une langue pivot avant leur traductions suivantes). La comparaison des variantes de traductions ainsi obtenues doit permettre de proposer des hypothèses pour l'identification des segments porteurs d'écarts.
- d'examiner ensuite :
  1. les différentes formes d'écart en traduction, cf plus supra, et leur impact sur la qualité de la traduction,
  2. les approches et les limites de l'intelligence artificielle en traduction, en particulier les modèles de traduction basés sur l'apprentissage automatique (de type Transformer comme GPT, etc.), et leur capacité à gérer les écarts,
  3. les innovations possibles, telles que l'intégration de modèles hybrides alliant IA et interventions humaines, pour surmonter ces limites. Et comment améliorer la prise en charge de ces écarts et ouvrir la voie à des systèmes de traduction plus en phase avec le défi de la traduction littéraire.

Tous ces travaux s'inscriront dans la démarche engagée conjointement par Avignon Université et la Cité Internationale de la Langue Française autour du lien Traduction et Patrimoine en relation avec le projet d'inscription de « La traduction au patrimoine immatériel de l'humanité de l'Unesco» [Lombard, 2024]. Cette démarche unique représente une solution pour répondre à une inquiétude très vive de tous les acteurs de la traduction face à l'émergence de plus en plus incontrôlable de l'IA dans leurs professions et des risques et dangers associés (voir par exemple l'initiative IA et traduction littéraire : les traductrices et traducteurs exigent la transparence, portée par l'Association pour la promotion de la traduction littéraire). L'enjeu ne se limite bien sûr pas à la disparition de l'activité humaine en traduction mais aussi sur les effets produits par la disparition de l'acte créatif porté par les traducteurs dans le processus et qui est un garant de la continuité de l'œuvre initiale et son inscription dans un patrimoine culturel de connaissances universelles.

Enfin, le département numérique du ministère de la culture a confirmé aux encadrants que dans le cadre de ces travaux ils sont intéressés par notre participation sur le volet de la traduction, aux discussions autour des évolutions du programme Compar:IA, comparateur public d'IA conversationnelles.

2- Profil du/de la candidat(e)

« Candidates for this position should have a master's degree in computer science, computational linguistics, or a related discipline. Experience with machine learning including deep learning and large language models is desired. The research will be conducted in French or English. »

- Étudiant possédant un Master 2 ou Diplôme d'école d'ingénieur en informatique, Spécialités Computational Linguistics, Traitement du Langage Naturel (NLP), IA ou Machine Learning
- Solides compétences en ingénierie informatique (Python, Linux, shell, git)
- Expérience attendue en Computational Linguistics, NLP et en Machine Learning, Deep Learning
- Connaissance des grands modèles de langage (LLM), ou IA générative en général
- Langues : anglais académique, la connaissance du français et/ou la maîtrise d'autres langues sera considérée comme un plus

3- Opportunités de mobilité à l'international du/de la doctorant(e) dans le cadre de sa thèse

La phase de diagnostic de la première année de thèse devrait préciser les besoins en collaboration les plus importants pour la réalisation des travaux. D'emblée les encadrants sont en relation de collaboration avec quelques universités dans lesquelles des échanges pourraient permettre de favoriser le développement de certaines parties de la thèse :

- l'université de Cagliari
- l'université Degli Studi de Milan
- l'université Aristote de Thessalonique
- l'ADAPT Centre de Trinity College à Dublin
- l'IDIAP de Martigny, Suisse

4- Références bibliographiques

Laurent Lombard, « Les écarts créateurs ou le défaut comme positivité dans l'art de traduire », in *Le* défaut, Revue Silène, Paris, Université de Paris Nanterre, 2023.

Laurent Lombard, « La traduction est un art, un patrimoine immatériel », In Les univers du livre – Actualités, 2024 https://actualitte.com/article/120561/edition/la-traduction-est-un-art-un-patrimoine-immateriel

Jean-Yves Masson « De la traduction comme acte créateur : raisons et déraisons d'un déni ». Meta 62, no 3, 2017, 635–646. https://doi.org/10.7202/1043954ar

Yves Chevrel, « Les traductions: un patrimoine littéraire? Revue d'histoire littéraire de la France », no 97(3), 1997, 355-360. https://doi.org/10.3917/rhlf.g1997.97n3.0355.

Raphaël Rubino, Stéphane Huet, Fabrice Lefèvre, and Georges Linarès, « Post-édition statistique pour l'adaptation aux domaines de spécialité en traduction automatique (Statistical Post-Editing of Machine Translation for Domain Adaptation) [in French] », In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN, pages 527–534, Grenoble, France. ATALA/AFCP, 2012.

A. Njifenjou, V. Sucal, B. Jabaian and F. Lefèvre, « Open-Source Large Language Models as Multilingual Crowdworkers: Synthesizing Open-Domain Dialogues », In Several Languages With No Examples in Targets and No Machine Translation », In 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics, Albuquerque (NM), 2025

A. Njifenjou, V. Sucal, B. Jabaian and F. Lefèvre, « Language Portability Strategies for Open- domain Dialogue with Pre-trained Language Models from High to Low Resource Languages », In The 13th International Workshop on Spoken Dialogue Systems Technology (IWSDS '23), Los Angeles (CA), 2023

Njifenjou, A., Sucal, V., Jabaian, B., and Lefèvre, « Portabilité linguistique des modèles de langage pré-appris appliqués à la tâche de dialogue humain-machine en français », In 18e Conférence en Recherche d'Information et Applications--16e Rencontres Jeunes Chercheurs en RI--30e Conférence sur le Traitement Automatique des Langues Naturelles--25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (pp. 148-158), Paris, 2023a.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia (2017). "Attention is All you Need". Advances in Neural Information Processing Systems. 30

☒ **J'ai informé le Directeur de mon unité du dépôt de cette proposition de sujet de thèse**

Les sujets devront être adressés avant le 9 décembre 2024 midi aux adresses fr-agorantic@univ-avignon.fr et intermedius@univ-avignon.fr

Maximum 5 pages (titre fichier : acronymesujet-labo-nom-Dirthèse-AGORANTIC.InterMEDIUS)