DOI: https://doi.org/10.17234/SRAZ.70.5

UDC: 81'25
UDC: 81'322.4
Original scientific paper
Reçu le 11 septembre 2025
Accepté pour la publication le 10 novembre 2025

La machine est-elle en train de remplacer l'homme ? Évaluation de traductions spécialisées réalisées par l'intelligence artificielle

Klara Miholić
Marta Richter
Faculté de philosophie et lettres
Université de Zagreb
miholicklara@gmail.com
mpetrak@ffzg.hr

Les avancées récentes en technologie générative ont eu un impact considérable sur l'industrie langagière. Si la traduction automatique (TA) est devenue la règle au fil des ans, atteignant un haut degré de performance suite à l'introduction des modèles neuronaux (la TAN), l'apparition de grands modèles de langage (LLMs) a introduit une vraie nouveauté dans le domaine. Dans le cadre de cet article, nous explorons l'emploi de ChatGPT, l'un des outils les plus populaires d'IA générative, dans la traduction de textes spécialisés. Si de nombreuses études confirment que la qualité des traductions produites par l'IA s'améliore rapidement, il semble que cela n'est pas encore vrai pour les textes spécialisés, qui reposent sur une terminologie précise. C'est pourquoi nous avons entrepris, dans cet article, d'analyser la qualité des traductions spécialisées générées par ChatGPT dans les domaines de la médecine, de l'économie et du droit sur la base d'une comparaison avec des traductions humaines. Des métriques automatiques ainsi que l'évaluation humaine ont été utilisées pour l'analyse. D'après nos résultats, les rendements de ChatGPT nécessitent une phase obligatoire de post-édition, et contiennent généralement le plus grand nombre d'erreurs lexicales. C'est le domaine juridique qui s'est avéré être le plus difficile pour l'IA. Nous avons conclu que ChatGPT peut être utile dans une certaine mesure pour faciliter et accélérer la traduction des textes spécialisés du français en croate, mais il ne peut toujours pas remplacer l'expertise et la précision humaines.

Mots-clés : traduction automatique (TA), intelligence artificielle (IA), technologie générative, grands modèles de langage (LLMs), ChatGPT, évaluation humaine et automatique de la TA, textes spécialisés

1. Tour d'horizon de l'emploi des technologies dans le secteur de la traduction

Le domaine de la traduction est depuis des décennies inextricablement liée à l'évolution de la technologie. Le quotidien des traducteurs est largement fondé non seulement sur l'emploi de la TAO (traduction assistée par ordinateur), s'appuyant sur les mémoires de traduction – faites par les humains – mais aussi sur celui de la TA (traduction automatique), produite uniquement par ordinateur, sans intervention humaine. C'est la TA qui est au centre de notre intérêt dans cet article. Ce type de traduction a vu une évolution fulgurante, s'étalant sur une série de trois générations de systèmes fondés sur : 1) les règles, 2) les probabilités statistiques et, finalement, 3) les modèles neuronaux, toujours d'actualité. Les progrès techniques ont été accompagnés de résultats toujours meilleurs, ainsi que d'un nombre croissant de langues traitées[1]. Désormais, certains fournisseurs de services de traduction automatique atteignent un niveau de performance similaire à l'humain (par ex. Yan et al. 2024).

Alors que l'industrie langagière, dont fait partie le secteur de la traduction, témoigne depuis de longues années de l'influence croissante des TAO et TA sur la vie des traducteurs, l'apparition de grands modèles de langage, tels que ChatGPT, a provoqué une vraie onde de choc dans le secteur. C'est en janvier 2023, avec le lancement de ChatGPT, que les grands modèles de langage ont atteint le grand public, avec une interface qui rend possible une « communication » assez naturelle avec l'intelligence artificielle (IA). Le phénomène a eu pour résultat la montée sans précédent de l'utilisation de cette nouvelle technologie[2]. Deux ans après, l'IA fait partie de notre vie quotidienne, et ne cesse d'attirer l'intérêt public.

De façon similaire, l'IA transforme le secteur de la traduction. Du côté des professionnels, l'IA génère les mêmes réactions que la traduction automatique il y a quelques années, selon les données de ELIS[3] pour 2025 (2025 : 40). Bien que l'utilisation réelle de l'IA ait été très limitée en 2024 (autour de 10 %) auprès des prestataires de services linguistiques et des traducteurs indépendants en Europe, cela pourrait changer radicalement en quelques années seulement, comme l'a montré l'utilisation de la TA dans le passé (*ibid*.). Les professionnels identifient l'IA et la TA comme les principales tendances dans le secteur, tant dans le sens

[1] Google Traduction, le service de traduction automatique de Google, l'un des plus populaires à l'échelle mondiale, a rajouté 110 nouvelles langues en juin 2024 ; voir : https://blog.google/products/translate/google-translate-new-languages-2024/. Ainsi, le nombre de langues que le modèle prend en charge s'est élevé à 249, selon certaines données (https://en.wikipedia.org/wiki/Google_Translate, accès en septembre 2025).

[2] On évoque plus de 700 millions d'utilisateurs actifs hebdomadaires, soit environ 10% de la population adulte mondiale, et 18 milliards de messages échangés chaque semaine en 2025 (https://www.usine-digitale.fr/article/chatgpt-entre-les-mains-du-grand-public-ca-donne-quoi.N2237822).

[3] ELIS (*European language industry survey*) est une enquête annuelle qui analyse les tendances dans l'industrie langagière à l'échelle européenne. Disponible sur https://elis-survey.org/wp-content/uploads/2025/03/ELIS-2025_Report.pdf (accès en avril 2025).

positif que négatif. Compte tenu des tendances générales dans le secteur de la traduction mentionnées ci-dessus, l'objectif principal de cet article est d'évaluer la qualité des traductions réalisées par ChatGPT du français en croate. Pour faire cela, nous allons nous concentrer sur les textes spécialisés car ceux-ci restent peu explorés dans le domaine de la TA produite à l'aide de l'IA (Jiang / Zhang 2024 : 4).

2. Intelligence artificielle (IA) et traduction automatique (TA) : quels rapports ?

Comme nous avons déjà expliqué, la traduction automatique consiste à traduire des textes par ordinateur, sans aucune intervention humaine. Selon l'enquête ELIS pour 2025, plus de 50 % des professionnels indépendants et des sociétés de services linguistiques utilisent cette technologie (2024 : 5). De nombreuses approches et méthodes sont apparues au cours de l'évolution de la TA comme nous avons indiqué dans l'introduction, chacune ayant ses propres avantages et limites. La phase actuelle du développement de la TA, appelée aussi *traduction automatique neuronale* (TAN), a marqué une avancée majeure en produisant des traductions d'une qualité assez élevée, proche de celles réalisées par un humain (Barbin 2023 : 51). Grâce aux réseaux de neurones artificiels, ces systèmes évoluent en temps réel en intégrant de nouveaux segments, mais leur performance dépend de l'apport constant de données bilingues de qualité (Barbin 2023 : 51).

On peut parler aujourd'hui d'une véritable *démocratisation* de la TA, qui s'est imposée au grand public, et ce notamment grâce à des sites Internet qui proposent des traduction rapides et gratuites dans de nombreuses langues (Champsaur 2013 : 21). Cette démocratisation concerne aussi le fait qu'actuellement de nombreux textes qui n'auraient jamais été traduits auparavant (blogs, commentaires sur les réseaux sociaux, courriels échangés avec des personnes parlant d'autres langues, etc.) peuvent l'être à présent, facilement et en quelques instants à peine.

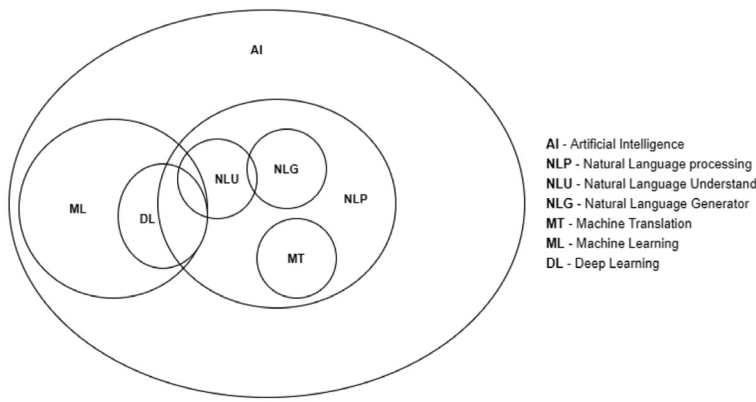
Il s'agit pourtant d'une technologie à employer avec prudence, surtout si un texte traduit à l'aide de la TA est destiné à être publié et / ou utilisé dans un contexte professionnel (O'Brien 2022 : 105). Plus précisément, les rendements de la TA devraient toujours être accompagnés d'une phase de relecture par un traducteur, appelée *post-édition*. Cette combinaison de TA et de post-édition peut réduire considérablement le travail du traducteur (O'Brien 2022 : 105 ; Robert : 2010). Il faut savoir qu'aujourd'hui, dans le secteur de la traduction, la TA est normalement utilisée ensemble avec la TAO[4], c'est-à-dire avec les mémoires de traduction.

[4] Témoignent de cette symbiose des logiciels tels que Trados, outil de traduction commercial le plus utilisé dans le monde, qui est fondé sur une combinaison des TAO et TA.

2.1. Grands modèles de langage

L'intelligence artificielle est la capacité d'un ordinateur – une machine – à effectuer des tâches généralement associées aux processus intellectuels caractéristiques du cerveau humain, comme la capacité de raisonner, de résoudre des problèmes ou de générer du langage (Silva / Fonseca 2019). L'un des domaines importants de l'IA est le traitement automatique des langues naturelles (TALN). Le TALN est un champ interdisciplinaire qui combine linguistique computationnelle, science informatique, science cognitive et intelligence artificielle (Deng / Liu 2018 : 1). Les applications typiques du TALN incluent reconnaissance vocale, analyse syntaxique, recherche d'informations, analyse des sentiments, ainsi que traduction automatique (*ibid*.). C'est sur ce dernier point que le TALN rejoint les intérêts de cet article.

Il ressort de ce qui précède que l'avènement de modèles tels que ChatGPT n'était pas le point d'entrée de l'IA dans le secteur de la traduction ; au contraire, elle y est présente depuis plusieurs décennies. Les relations entre l'IA et la TA seront plus claires si l'on regarde le schéma ci-dessous, proposé par Šuman (2021 : 372) :



Graphique 1 : Les domaines qu'englobe l'intelligence artificielle (IA)

L'on y voit bien que l'IA englobe deux grands ensembles : apprentissage automatique (ang. machine learning (ML) et TALN (ang. natural language processing, NLP). Ce dernier inclut entre autres, la traduction automatique (ang. machine translation (MT)), qui nous concerne en particulier dans cet article. S'ensuit du graphique aussi qu'il existe des points communs entre ces deux grands ensembles, comme nous avons souligné précédemment.

Aujourd'hui on entend surtout parler de l'intelligence artificielle *générative*. Il s'agit d'un domaine de l'apprentissage automatique dédié à la création de nouveaux contenus. Alors que les modèles d'apprentissage automatique sont principalement axés sur classification, prédiction et prise de décision, les modèles d'IA générative apprennent la structure des données d'entraînement et

les exploitent pour produire de nouvelles données, telles qu'images, musique, vidéos ou textes cohérents, qui ressemblent au texte produit par des humains (Lath et al. 2025).

Il faut ajouter aux notions expliquées précédemment un autre mot clé, celui des *grands modeles de langage*[5] (ang. large language models, LLMs). Les grands modèles de langage sont des systèmes d'IA conçus pour « comprendre » et générer du texte, coder, et bien plus encore (Ozdemir 2023). Ces modèles sont entraînés sur de grandes quantités de données textuelles. Au niveau linguistique, l'un de leurs plus grands avantages est le fait de prendre en compte un contexte plus large pour analyser comment un mot est lié à tous les autres mots dans un contexte donné. Des LLMs populaires modernes sont BERT, T5 et GPT, développés par Google et Open AI respectivement. Ils sont facilement accessibles au public grâce à des interfaces telles que Copilot de Microsoft, Llama de Meta ou Gemini de Google, de même que l'incontournable ChatGPT d'Open AI. Grace à leur succès, on appelle les années 2020 « l'ère des LLMs » (Kamath et al. 2024 : 6).

Le grand modèle de langage qui est au cœur de notre intérêt dans cet article est ChatGPT. Basé sur l'architecture GPT (angl. *Generative Pre-Trained Transformer*), il a été développé en 2018 par l'entreprise américaine OpenAI (Roumeliotis / Tselikas 2023 : 2-3) et ouvert au public en décembre 2022. Il s'agit d'un chatbot Web formé pour répondre aux requêtes des utilisateurs (Kamath et al. 2024 : 1) qui a provoqué une sensation mondiale et connu la croissance la plus rapide de l'histoire[6] (*ibid*.). Comme les autres LLMs, ChatGPT n'a pas été conçu exclusivement pour la traduction, mais il a démontré une capacité technique suffisante pour produire des traductions fluides et cohérentes, rivalisant voire surpassant des services de traduction automatique comme Google Traduction et DeepL (Lee 2023 : 2). C'est un modèle linguistique très intelligent, capable d'améliorer l'apprentissage grâce aux retours fournis par les utilisateurs ou l'entraînement sur de nouvelles données. (*ibid.* : 4).

Contrairement aux outils de TA qui produisent un texte entier sans intervention humaine, ChatGPT a besoin d'une instruction ou d'un message-guide pour générer les réponses souhaitées. Ces instructions, appelées *prompts* (de l'anglais *prompts*), *invites* ou *amorces*, jouent un rôle essentiel dans l'optimisation des résultats. Un prompt bien conçu peut améliorer la qualité et la pertinence des résultats de ChatGPT de manière significative, tandis qu'un prompt mal conçu peut conduire à des réponses insatisfaisantes ou erronées (Ekin 2023 : 3).

---

[5] Certains auteurs les appellent *giga modèles de langue* (par ex. Wisnievski 2025). Les auteurs écrivant en français ont tendance à garder l'acronyme anglais LLMs, ce que nous avons décidé de faire aussi.

[6] Elle a attiré plus de 100 millions d'utilisateurs mensuels avant janvier 2023 (Kamath et al. 2024 : 1).

3. Objectif et hypothèses

Le but de cet article est d'analyser la performance de l'intelligence artificielle dans le domaine de la traduction spécialisée du français vers le croate. Il s'agit, à notre connaissance, de la première étude portant sur cette paire linguistique, combinant une langue mondiale (le français) avec une langue plus petite (le croate), disposant de moins de ressources[7]. L'analyse repose sur une comparaison des traductions réalisées par ChatGPT avec des traductions humaines, et ce à l'aide de méthodes d'évaluation humaine et automatique.

Nous avons formulé trois hypothèses :

1) Les traductions générées par ChatGPT n'atteignent pas le niveau de qualité d'une traduction humaine.
2) Les textes dans le domaine du droit seront de la qualité plus basse par rapport aux textes des domaines de la médicine et de l'économie.
3) Le plus grand nombre d'erreurs dans la traduction de ChatGPT proviendra de la catégorie des erreurs lexicales.

Avec la première hypothèse, nous cherchons à déterminer si ChatGPT est capable de générer des traductions nécessitant peu de modifications par le post-éditeur. Il sera également intéressant d'observer les performances de ChatGPT dans une langue plus petite comme le croate parce que la plupart des recherches menées jusqu'à présent ont confirmé une grande aptitude des LLM à traduire des langues bien dotées en ressources, comme l'anglais ou l'allemand, alors que les études portant sur des langues peu dotées restent sous-représentées (Jiang / Zhang 2024 : 4). De même, Jiao et al. (2023 : 3) ont constaté que les résultats de ChatGPT sont comparables à ceux des systèmes de TAN, tels que Google Traduction, dans le cas des grandes langues européennes, mais qu'il éprouve des difficultés avec les langues peu dotées ou typologiquement très différentes. Dans ce contexte, les études portant sur la TA relative à la paire de langues français – croate ont été, à notre connaissance, jusqu'à présent à peine explorées.[8]

La deuxième hypothèse découle de l'idée qu'il existe une corrélation entre le domaine et la qualité de la traduction. Plusieurs études soulignent l'importance d'évaluer les performances des systèmes de traduction automatique dans des domaines spécialisés car les résultats peuvent varier selon le type de contenu traité, qu'il s'agisse de textes scientifiques, juridiques ou autres (Son / Kim 2023 : 16).

La troisième hypothèse repose sur l'idée que la nature des textes choisis (ici : des textes spécialisés) entraînera un taux plus élevé d'erreurs lexicales, principalement en raison de l'utilisation incorrecte de termes spécialisés.

---

[7] Dans le contexte du traitement du langage naturel, le croate est considéré comme une langue dotée de peu de resources (par ex. Filipović Petrović et al. 2024) ou bien comme une langue disposant d'un niveau moyen de ressources (« a moderately resourced language », Štefanec et al. 2024).

[8] Il existe un article scientifique traitant de l'application de la TA dans le domaine littéraire (Petrak, Uremović et Pavelin Lešić 2022) et quelques mémoires de Master rédigés à la Chaire de langue française de la Faculté de philosophie et lettres de Zagreb.

4. Traduction des textes spécialisés dans le contexte de la TA

Dans le cadre de cette étude, nous avons analysé des textes juridiques, économiques et médicaux. Il faut savoir que les systèmes de TA restent généralement peu performants dans les domaines spécialisés et dans les langues peu dotées en ressources (Bajčić / Golenko 2024 : 172). Bajčić et Golenko (2024 : 171) soutiennent que la traduction spécialisée reste problématique pour l'automatisation, notamment au niveau de la terminologie.

De nombreux auteurs soulignent que la traduction des textes juridiques est l'un des domaines les plus complexes de la traduction spécialisée (Gémar 2015 ; Wiesmann 2019 ; Janigová 2025). La traduction juridique diffère de la traduction de domaines tels que la chimie ou la médecine, par exemple, en ce sens que le droit est avant tout un phénomène national entretenant des liens forts avec la société et la tradition desquelles il provient (cf. Šarčević 1997 ; Biel 2022). C'est pourquoi le principal défi pour les traducteurs juridiques réside dans la divergence des systèmes juridiques (Šarčević 1997 : 13). En comparaison, des domaines tels que la médecine ou l'économie sont souvent plus standardisés (Biel / Sosoni 2017 : 353). Dans le domaine juridique, la TA peut être un outil précieux dans le processus de traduction, mais celle-ci doit toujours être accompagnée d'une post-édition car l'expertise humaine reste indispensable pour assurer la qualité (Omazić / Šoštarić 2023).

Les terminologies économique et financière possèdent une dimension culturelle, influencée par les divergences historiques et idéologiques entre les systèmes économiques (Biel / Sosoni 2022 : 352). Cependant, la mondialisation et l'internationalisation des pratiques commerciales ont favorisé une certaine harmonisation, rendant la terminologie des affaires plus universelle que celle du droit. La domination des économies anglophones et le statut de l'anglais en tant que *lingua franca* des affaires ont conduit à une large adoption et intégration de termes anglais dans de nombreuses langues (*Ibid*. : 352-353). Selon certaines études, la TA dans le domaine des textes économiques et financiers n'atteint pas la qualité de la traduction humaine (Boumparis / Giannoutsos 2023) ; toutefois, des modèles entrainés pour des tâches spécifiques surpassent systématiquement la performance des LLMs dans tous les groupes de langues européennes (Oncevay et al. 2025).

Quant au domaine médical, celui-ci a été largement influencé par les mots grecs et latins, donnant lieu à un grand nombre d'internationalismes. Selon des études existantes, les systèmes de TA à usage général présentent des résultats limités sur les textes médicaux, alors que des systèmes entraînés sur des corpus spécialisés montrent de bien meilleures performances (Renato et al. 2018).

5. Méthodes d'évaluation de la traduction automatique

De nombreuses études ont souligné l'importance d'évaluer la qualité de la TA. S'il n'existe toujours pas de définition universellement acceptée de la qualité en TA (Rossi / Carré 2022 : 52), celle-ci est souvent évaluée en termes de précision,

de fluidité ou de proximité avec la traduction humaine (Jiang / Zhang 2024 : 3). La qualité des rendements de la TA peut être évaluée grâce aux méthodes automatiques et à l'évaluation humaine, chacune présentant des avantages et des inconvénients (Rossi / Carré 2022 : 53).

Les métriques automatiques sont largement utilisées dans la recherche en raison de leurs rapidité, efficacité, fiabilité et moindre coût[9] (Jiang / Zhang 2024 : 3). Parmi celles-ci, on distingue les mesures qui évaluent le degré de concordance entre une traduction automatique et une traduction humaine de référence (« étalon-or », en angl. *gold standard*) en utilisant des comparaisons linguistiques et statistiques. Il y a des métriques qui opèrent au niveau superficiel de la phrase (Nakhlé 2023 : 147), dont TER et BLEU[10], que nous emploierons dans cette étude. En outre, Nakhlé distingue (*ibid.* : 148) les *métriques apprises*, aussi appelées *métriques neuronales*, qui reposent sur des modèles de langage pré-entraînés et utilisent l'apprentissage automatique. Nous utiliserons BERTScore et COMET[11] comme exemples de ces métriques.

Si toutes ces mesures permettent de quantifier la qualité de la TA, elles ne parviennent pas à en saisir toute la complexité (Son / Kim 2023 : 5). Par exemple, les traducteurs humains accordent plus d'attention à des dimensions telles que le public visé, la sensibilité culturelle, le respect des normes de la traduction, la cohérence textuelle et l'aspect pratique (Jiang / Zhang 2024 : 1). Il est donc essentiel de combiner les métriques automatisées avec l'évaluation humaine pour obtenir une évaluation plus complète (Son / Kim 2023 : 5).

Une autre méthode d'évaluation consiste à compter les erreurs de traduction sur la base d'une typologie prédéfinie (*Ibid*. : 58). C'est justement la méthode que nous emploierons pour effectuer l'évaluation humaine, en nous appuyant sur la typologie proposée par Pavlović (2016), avec quelques modifications mineures, car celle-ci a été conçue en tenant compte des spécificités de la langue croate.

6. Méthodologie

Nous avons choisi deux textes par domaine de spécialité, chacun comportant environ 6,000 caractères, espaces compris. Les traductions ont été réalisées par une étudiante dans le cadre de son mémoire de Master[12] et corrigées par une traductrice professionnelle experimentée. Ces traductions corrigées ont joué le rôle de versions de référence (« étalon-or ») et ont ensuite été employées pour la comparaison avec

[9] L'évaluation humaine implique un processus plus long pour lequel le traducteur devrait normalement être rémunéré (Rossi / Carré 2022 : 53).

[10] Pour plus de détails sur ces deux métriques, voir Snover et al. (2006) et Papineni et al. (2002).

[11] Pour plus de détails concernant BERTscore et COMET, voir Zhang et al. (2020) et Rei et al. (2020).

[12] Les données de l'analyse présentée dans cet article sont tirées du mémoire de Master intitulé *Évaluation de la qualité de traductions spécialisées produites par ChatGPT*, rédigé par Klara Miholić et soutenu à l'Université de Zagreb en juillet 2025.

les traductions générées par ChatGPT. Il faut souligner que nous avons utilisé GPT-4o[13], l'une des versions payantes du service, introduite en mai 2024[14].

Il faut préciser également l'approche que nous avons employée au regard des prompts. À ce jour, un grand nombre d'études ont montré que la qualité de la traduction est influencée par les prompts utilisés[15] (Lee 2023 : 4 ; Jiang / Zhang 2024 : 2). ChatGPT ayant été entraîné sur un grand nombre de données multilingues d'usage général, la spécification du domaine peut contribuer à garantir que le texte généré soit davantage basé sur les données fournies, plutôt qu'influencé par ses connaissances générales préalables (Gao et al. 2023 : 2-3). Tenant compte de cela, nous avons utilisé le prompt suivant, formulé en croate :

> « Le texte suivant est un texte spécialisé écrit en français dans le domaine [de l'économie / de la médecine / du droit]. Traduis-le en croate en utilisant la terminologie et le style appropriés. »

6.1. Évaluation automatique et humaine

Pour évaluer les traductions produites par ChatGPT, nous avons utilisé le site web MATEO[16], qui combine jusqu'à six métriques automatiques et, grâce à son interface simple, permet une utilisation rapide et facile des métriques sélectionnées. Pour l'évaluation humaine, comme mentionné *supra*, nous avons employé le modèle de Pavlović (2016 : 285), selon lequel les erreurs sont réparties dans les catégories suivantes : A. orthographe, B. lexique, C. morphosyntaxe et D. autres erreurs, avec des sous-catégories (v. Annexe). Nous avons légèrement adapté la classification de Pavlović afin qu'elle corresponde mieux aux particularités de la traduction utilisant la technologie générative. Plus précisément, nous avons incorporé certains éléments de la typologie d'erreurs MQM[17], en renommant la catégorie D « précision » et en y ajoutant la sous-

---

[13] OpenAI. "Hello GPT-4" https://openai.com/index/hello-gpt-4o/ (accès en avril 2025)

[14] Il faut souligner que l'entreprise OpenAI améliore constamment ses modèles GPT. En ce moment, il en existe plusieurs, chacun proposant des fonctionnalités adaptées à des tâches spécifiques. Pour faciliter l'utilisation du modèle GPT-4o, nous avons souscrit à ChatGPT Plus, qui offre un accès plus stable au système s'il est utilisé par un grand nombre d'utilisateurs, des temps de réponse plus rapides et un accès prioritaire aux nouvelles fonctionnalités. Voir : OpenAI. "What is ChatGPT Plus?" https://help.openai.com/en/articles/6950777-what-is-chatgpt-plus#h_d78bb59065 (accès en avril 2025)

[15] Ceux-ci sont aussi importants qu'il est en train de se développer une discipline appelée en anglais *prompt engineering*, ou *ingénierie des prompts*, qui désigne « la procédure de conception, d'amélioration et d'optimisation des instructions en entrée (prompts) d'un modèle d'IA conversationnel et génératif, lui permettant de produire des sorties précises, pertinentes et utiles » (De Sousa Cardoso / Parise 2023).

[16] MATEO: MAchine Translation Evaluation Online https://mateo.ivdnt.org/ (accès en mai 2025)

[17] MQM (Multidimensional Quality Metrics) https://themqm.org/the-mqm-full-typology/ (accès en mai 2025)

catégorie D.d. Nous avons considéré que les erreurs de la catégorie « précision » (ou exactitude) surviennent lorsque le contenu cible ne reflète pas fidèlement le message du texte source en raison d'une distorsion, d'une omission ou d'un ajout au message. Nous avons également ajouté une nouvelle catégorie relative au style (catégorie E.). Ces erreurs se réfèrent aux solutions jugées acceptables d'un point de vue grammatical, mais inappropriées du point de vue de style et / ou registre linguistique inadéquat.

Comme chez Pavlović (2016 : 284), toute partie de la traduction que le post-éditeur n'a pas jugée comme prête pour publication sans correction a été considérée comme une erreur. Au cours du processus d'évaluation, nous avons tenu compte de l'objectif de chaque traduction, du public cible, du type de texte et du domaine de spécialisation. Si une erreur pouvait être classée dans plus d'une catégorie, elle était comptée dans chacune de ces catégories. Pour une analyse des données plus facile, le nombre des erreurs a ensuite été exprimé en pourcentages.

7. Résultats et analyse des évaluations

Dans ce chapitre, nous présenterons l'analyse de l'évaluation automatique et humaine, en commençant par les données statistiques obtenues par l'évaluation automatique.

7.1. Évaluation automatique des traductions automatiques

Les résultats des quatre métriques utilisées (BERTScore, COMET, BLEU et TER) sont présentés au Tableau 2 ci-dessous.

	BERTScore	**COMET**	**BLEU**	**TER**
M1	86,81	93,86	25,59	79,94
M2	85,49	92,48	24,84	63,72
Moyenne	86,15	**93,17**	25,215	71,83
E1	87,71	92,63	33,38	55,36
E2	87,94	93,13	31,93	61,32
Moyenne	**87,825**	92,88	**32,655**	**58,34**
D1	80,73	91,09	17,88	72,03
D2	83,04	92,4	30,38	65,12
Moyenne	81,885	91,745	24,13	68,575

Tableau 2 : Les résultats des métriques automatiques
(M1, M2 – médicine ; E1, E2 – économie ; D1, D2 – droit)

Rappelons que BERTScore et COMET, en tant que des métriques neuronales, évaluent l'aspect sémantique, et un score plus élevé indique une meilleure qualité de rendement. Il en va de même pour BLEU, alors que dans le cas de TER,

qui identifie le nombre d'éditions, un score plus faible désigne une meilleure traduction. Le score maximal pour chaque métrique est de 100.

Il convient de noter que pour la métrique BLEU, les scores entre 10 et 19 suggèrent que le sens général du texte est difficile à comprendre, tandis que les scores entre 20 et 29 indiquent que le sens est compréhensible mais qu'il y a encore de nombreuses erreurs dans la traduction. Les valeurs entre 30 et 39 indiquent une traduction généralement compréhensible et de meilleure qualité, alors qu'en général, les scores supérieurs à 50, très rares, signalent une qualité très élevée, proche de la production humaine (Omazić / Šoštarić 2023 : 79).

D'après le tableau, nous pouvons conclure que les textes du domaine de l'économie ont obtenu les meilleurs résultats selon trois des quatre métriques (chiffres en gras). Le score de COMET est légèrement plus élevé pour les textes médicaux, mais il s'agit d'une petite différence. Nous pouvons donc conclure que, selon les métriques automatiques, les traductions des textes économiques sont de meilleure qualité que celles des deux autres domaines.

En outre, notre hypothèse selon laquelle les textes du domaine juridique seraient de qualité inférieure a été confirmé. Comme le montre le tableau (cases grises), les textes de ce domaine obtiennent en moyenne les scores les plus bas sur trois des quatre métriques utilisées. Ils n'arrivent qu'en deuxième position pour la métrique TER ; les textes du domaine médical ont des résultats légèrement plus élevés. Cependant, étant donné que les trois autres mesures placent les textes du droit en dernière position, nous pouvons conclure que ce domaine se distingue par la plus faible qualité des traductions.

L'analyse des résultats individuels montre que le texte D1 (sur les obligations) a obtenu les scores les plus bas sur BERTScore (80,73) et COMET (91,09), alors que le score BLEU est particulièrement peu élevé par rapport à ceux des autres textes (17,88). En outre, son score TER est le deuxième plus élevé, dépassé uniquement par le texte M1. Ces résultats indiquent clairement que la TA du texte D1 est la traduction de moins bonne qualité de notre corpus, ce qui n'est pas surprenant étant donné sa complexité et les différences existant entre les systèmes juridiques croate et québécois. De plus, contrairement à l'article scientifique sur le système pénitentiaire (D2), il s'agit ici d'un texte normatif (législatif) dont la syntaxe et les constructions (collocations) sont très différentes de celles du discours scientifique, qui ne diffère pas significativement de la langue commune.

Selon les métriques BLEU et TER, on peut dire que la qualité des traductions générées par ChatGPT n'est pas satisfaisante et que leur utilité serait limitée. Pourtant, les valeurs élevées de BERTScore et COMET indiquent que les traductions automatiques sont assez réussies. Jiang et Zhang (2024 : 8) expliquent que les traductions produites par ChatGPT s'éloignent davantage des traductions de référence au niveau de la correspondance exacte des n-grammes[18], mais

[18] En traitement du langage naturel, un n-gramme désigne « un ensemble de n éléments successifs dans un document texte pouvant comprendre des mots, des nombres, des symboles, etc. » (voir par ex. https://fr.mathworks.com/discovery/natural-language-processing.html).

qu'elles sont tout de même fidèles aux références sur le plan sémantique grâce à l'architecture de ChatGPT qui lui permet de prendre en compte le sens général. Nos résultats confirment donc que les LLMs ont une grande capacité à produire des traductions cohérentes et fluides grâce à l'identification des relations sémantiques et du contexte global (*ibid*.).

7.2. Évaluation humaine

L'évaluation humaine a consisté à compter les erreurs selon 5 catégories (A. orthographe, B. lexique, C. morphosyntaxe, D. précision et E. style). Le Tableau 3 présente les résultats exprimés en pourcentage[19].

	M1	M2	M1+M2	E1	E2	E1+E2	D1	D2	D1+D2
A.a.	1,6	1,45	3,05	3,68	1,15	4,83	0,45	1,03	1,48
A.b.	0,2	/	0,2	0,31	0,29	0,6	/	/	0
A.c.	/	1,29	1,29	/	/	0	0,15	0,15	0,3
A-Total	1,8	2,74	4,54	3,99	1,44	5,43	0,6	1,18	1,78
B.a.	3,34	0,48	3,82	4,44	1,87	6,31	5,08	4,99	10,07
B.b.	2,23	5,79	8,02	3,83	3,44	7,27	8,82	2,94	11,76
B.c.	1,91	0,32	2,23	0,31	0,57	0,88	/	0,29	0,29
B-Total	7,48	6,59	14,07	8,58	5,88	14,46	13,9	8,22	22,12
C.a.	0,48	2,25	2,73	0,46	0,14	0,6	1,94	0,29	2,23
C.b.	0,96	0,32	1,28	0,61	0,14	0,75	0,45	0,15	0,6
C.c.	1,27	0,8	2,07	/	0,43	0,43	1,2	0,29	1,49
C-Total	2,71	3,37	6,08	1,07	0,71	1,78	3,59	0,73	4,32
D.a.	0,32	/	0,32	/	/	0	/	/	0
D.b.	2,07	2,41	4,48	2,76	1,43	4,19	6,58	0,44	7,02
D.c	0,64	1,61	2,25	0,31	0,29	0,6	0,45	/	0,45
D.d.	0,64	0,96	1,6	1,23	1,87	3,1	2,69	0,73	3,42
D.e.	/	1,13	1,13	0,31	3,3	3,61	0,9	0,29	1,19
D-Total	3,67	6,11	9,78	4,61	6,89	11,5	10,62	1,46	12,08
E-Total	2,55	1,77	4,32	1,68	3,44	5,12	4,48	4,41	8,89
TOTAL	18,21	20,58	38,79	19,93	18,36	38,29	33,19	16	**49,19**

Tableau 3. Les résultats de l'évaluation humaine selon les catégories d'erreurs exprimées en pourcentage (%) (M1, M2 – médicine ; E1, E2 – économie ; D1, D2 – droit)

[19] Les pourcentages ont été obtenus en divisant le nombre total d'erreurs pour chaque catégorie par le nombre total de mots dans le texte traité, qui a ensuite été multiplié par cent.

Les résultats de l'évaluation humaine des erreurs confirment une nouvelle fois notre hypothèse selon laquelle les textes du domaine juridique sont de la plus mauvaise qualité. Le pourcentage total d'erreurs pour ces textes est 49,19 %, contre 38,79 % pour la médecine et 38,29 % pour l'économie.

L'hypothèse selon laquelle la plupart des erreurs proviendraient de la catégorie lexicale (catégorie B) a également été confirmée. Les textes du domaine de la médecine contiennent 14,07 % d'erreurs lexicales, ceux du domaine de l'économie 14,46 % et ceux du domaine du droit 22,12 %.

Enfin, notre hypothèse générale a aussi été confirmée car les traductions générées par ChatGPT n'ont pas atteint la qualité d'une traduction humaine et auraient nécessité un certain effort de post-édition.

L'un des principaux inconvénients de ChatGPT est sa tendance à générer des réponses qui semblent claires et convaincantes mais qui peuvent être inexactes ou illogiques, c'est-à-dire des *hallucinations* (Waldo / Boussard 2024 : 1). Les LLMs produisent généralement des résultats fluides et cohérents, ce qui peut rendre les erreurs de traduction subtiles (par ex. mots omis, ajouts de mots inutiles, expressions inappropriées au contexte, etc.) plus difficiles à détecter. Dans la suite, nous présenterons l'analyse des erreurs par catégorie, en commençant par les catégories A (orthographe) et C (morphosyntaxe), qui concentrent le moins d'erreurs.

7.2.1. Fautes d'orthographe

Les fautes d'orthographe étaient principalement liées à l'usage incorrect de la ponctuation, notamment des virgules, des deux-points et des points-virgules, qui figuraient souvent dans la traduction sans adaptation à l'orthographe croate (A.a). Les erreurs liées aux lettres majuscules et minuscules étaient très rares (A.b), tout comme les autres types d'erreurs (A.c). Voici un exemple de cette catégorie :

Version originale	Traduction humaine	ChatGPT
Rien n'indique que les changements de la sensibilité des dérivés de l'artémisinine affectent un stade **asexué (**…)	Nema nikakvih naznaka da promjene u osjetljivosti derivata artemisinina utječu na **aseksualni** stadij (…)	Nema dokaza da promjene osjetljivosti utječu na druge **asexualne** stadije (…)

7.2.2. Morphosyntaxe

Dans cette catégorie, ChatGPT a eu des problèmes avec l'aspect verbal, un phénomène particulièrement complexe dans les langues slaves, dont le croate fait partie, ainsi qu'avec l'ordre des mots, qu'il laissait parfois comme dans l'original français, de même qu'avec la déclinaison de certains mots spécifiques, comme des sigles (CTA), ou bien avec l'emploi de la voix passive, qui est normalement évitée en croate, sauf dans certains discours (administratif, légal etc.). Un exemple de cette catégorie figure au tableau ci-dessous.

Version originale	Traduction humaine	ChatGPT
Actuellement six **CTA** sont recommandées par l'OMS…	SZO trenutačno preporučuje šest **ACT-ova**…	Trenutno je šest **CTA** preporučeno od strane SZO-a…

7.2.3. Erreurs lexicales

Dans les textes médicaux, l'abréviation CTA, mentionnée ci-dessus, représente un cas intéressant : en plus d'une erreur morphologique déjà expliquée, celle-ci a également été considérée comme un terme inapproprié. Dans les sources spécialisées croates, on emploie la forme anglaise du sigle (ACT), que l'on décline, alors que ChatGPT a simplement gardé la forme française (CTA). Un autre exemple d'erreurs lexicales figure ci-dessous.

Version originale	Traduction humaine	ChatGPT
Il peut être d'adhésion ou **de gré à gré**, **synallagmatique** ou **unilatéral**, **à titre onéreux** ou gratuit, commutatif ou aléatoire et **à exécution instantanée** ou **successive**; il peut aussi être **de consommation**.	Ugovor može biti adhezijski ili **tipski**, **dvostranoobvezan** ili **jednostranoobvezan**, **naplatan** ili besplatan, komutativan ili aleatoran te **s jednokratnom činidbom** ili **s uzastopnim činidbama**. Postoje i **potrošački ugovori**.	Može biti adhezijski ili **ugovoren po volji stranaka**, **dvostran** ili **jednostran**, **uz naknadu** ili besplatan, komutativan ili aleatoran, **s trenutnim** ili **sukcesivnim ispunjenjem**; može također biti **ugovor o potrošnji**.

Comme nous avons déjà souligné à plusieurs reprises, nos résultats ont montré que les textes juridiques comportent le plus grand nombre d'erreurs lexicales. Il y a une différence considérable entre nos traductions et celles de ChatGPT dans ce domaine, surtout dans le texte D1, qui est de nature normative. Alors que nous avons cherché à adapter les concepts et les termes juridiques au système croate, ChatGPT a produit des traductions plus littérales, nuisant à la précision terminologique et au style général. Le tableau ci-dessus montre à quel point il est difficile pour lui de maîtriser la terminologie juridique spécialisée, en particulier les noms de différents types de contrats. Et même lorsque la traduction est correcte (comme dans le cas de *komutativan* ou *aleatoran ugovor*), elle semble être le résultat d'une transposition littérale des termes français, dont les équivalents croates sont eux-mêmes très proches sur le plan lexical.

7.2.4. Précision

La deuxième catégorie des erreurs la plus fréquente était la catégorie D (précision), confirmant que les traductions produites par des modèles génératifs sont souvent marquées par omissions, ajouts ou contresens.

Version originale	Traduction humaine	ChatGPT
L'Oréal Luxe affiche une belle croissance **au premier semestre**, tirée par une progression **à deux chiffres** en Europe…	L'Oréal Luxe ostvario je značajan rast **u prvih šest mjeseci** zabilježivši **dvoznamenkasti** rast u Europi…	L'Oréal Luxe ostvaruje snažan rast, osobito u Europi…

Dans le tableau ci-dessus, il s'agit d'un cas d'omission. Dans la phrase en question, deux éléments ont été supprimés : *au premier semestre* et *à deux chiffres*, la dernière expression ayant été traduite par l'adjectif *snažan* ('fort'), une reformulation moins précise.

7.2.5. Style

Selon Jiang et Zhang (2024 : 9), le principal défi de ChatGPT réside précisément dans les erreurs liées au style, en particulier dans les tâches de traduction. Contrairement aux systèmes traditionnels de TAN, ChatGPT produit fréquemment des expressions maladroites ou excessivement idiomatiques, ce qui peut nuire à la fluidité et au ton naturel du texte. Ces problèmes stylistiques sont liés à la nature générative (« créative ») du modèle qui cherche avant tout à préserver le sens et la cohérence, parfois au détriment d'une formulation précise ou conventionnelle. Toutefois, l'apport de contexte permet de réduire sensiblement ces erreurs et d'améliorer la qualité stylistique globale des réponses (*Ibid.*).

Au tableau ci-dessous figure un exemple provenant du texte D2, qui a affiché un taux d'erreurs stylistiques légèrement plus élevé car il s'agit d'un article scientifique riche en terminologie spécialisée dont le style était plus libre, ce qui a parfois entraîné des traductions maladroites (en gras).

Version originale	Traduction humaine	ChatgGPT
La comparaison peut emprunter des **chemins déjà tracés** par la jurisprudence de la Cour de Strasbourg, **juridiction particulièrement dynamique**	U usporedbi mogu poslužiti već postojeći primjeri sudske prakse Suda u Strasbourgu, u čijoj se osobito dinamičnoj sudskoj praksi	Usporedba se može **osloniti na putove koje je već utabala** sudska praksa Suda u Strasbourgu, **suda poznatog po svojoj dinamici**,

7.2.6. ChatGPT au-delà des erreurs

Il existe un certain nombre d'exemples très réussis dans lesquels ChatGPT a simplifié le message ou trouvé des traductions élégantes sans nuire à la clarté ni au contenu du message. L'exemple ci-dessous prouve une fois de plus que

ChatGPT tient compte du contexte global du texte qu'il traduit. Au lieu de répéter les périodes de l'année mentionnées deux fois entre parenthèses, il a trouvé une expression élégante pour les remplacer (en utilisant l'expression *u istom razdoblju* 'sur la même période'). Le message est resté pareil et la phrase est devenue plus fluide. En revanche, nous avons gardé la construction de la version originale, qui est correcte, mais pas aussi réfléchie que celle de ChatGPT.

Version originale	Traduction humaine	ChatGPT
…la forte baisse du prix nominal des transactions (- 15,2 % entre le T2 2022 et le T2 2024) étant amplifiée par la hausse du déflateur du PIB (+ 10,1 % **entre le T2 2022 et le T2 2024**).	…pad realnih cijena jer je zbog rasta deflatora BDP-a (+10,1 % između drugog tromjesečja 2022. i drugog tromjesečja 2024.) dodatno pojačan pad nominalnih cijena transakcija (-15,2 % između drugog tromjesečja 2022. i drugog tromjesečja 2024.).	…nominalne cijene transakcija pale su za 15,2 % (od drugog tromjesečja 2022. do drugog tromjesečja 2024.), pri čemu je pad dodatno pojačan rastom deflatora BDP-a (+10,1 % **u istom razdoblju**).

8. Conclusion

Dans cet article nous avons exploré la qualité des traductions réalisées par un grande modèle de langage (LLM), ChatGPT. Il s'agit d'un type de l'intelligence générative qui est en train de transformer le secteur de la traduction. Afin d'analyser ses rendements, nous avons pris en compte trois domaines de spécialité qui figurent parmi les plus traduits : droit, médecine et économie. Selon la littérature existante, la traduction automatique n'obtient toujours pas des résultats satisfaisants dans le domaine des textes spécialisés, c'est pourquoi nous nous sommes concentrées sur ceux-ci. L'objectif principal était de comparer des traductions humaines avec celles produites par ChatGPT, afin de voir dans quelle mesure ChatGPT peut véritablement constituer un appui pour les traducteurs spécialisés. La paire linguistique choisie pour cette étude portait sur le français comme langue source et le croate comme langue cible.

Pour mesurer la qualité des traductions, nous avons employé quatre métriques automatiques (BERTScore, COMET, BLEU et TER) ensemble avec une évaluation humaine, qui consistait à classer les erreurs commises par ChatGPT selon cinq catégories (orthographe, lexique, morphosyntaxe, précision et style). Nos résultats ont montré que les traductions produites par ChatGPT n'ont pas atteint la qualité souhaitée, que la qualité des textes dans le domaine juridique est plus faible que dans les deux autres domaines, et que le plus grand nombre d'erreurs provient de la catégorie des erreurs lexicales.

Les résultats obtenus sont loin d'être représentatifs car ils ont été recueillis d'un corpus assez limité, mais ils ont néanmoins une valeur significative car ils se fondent sur une combinaison linguistique peu explorée dans le domaine de la TA (français-croate).

En conclusion, ChatGPT montre un réel potentiel pour faciliter et accélérer le travail des traducteurs. Cela dit, comme pour tout outil de TA, une relecture et une post-édition minutieuses sont indispensables, surtout lorsqu'il s'agit de textes qui ont vocation à être publiés. Il convient également de souligner que les performances de ChatGPT peuvent être améliorées de manière significative grâce à des prompts clairs et bien écrits. Des recherches ultérieures seraient nécessaires pour explorer davantage de combinaisons linguistiques, ainsi que pour étudier l'effet de différents types de prompts sur la qualité des traductions spécialisés.

8. Bibliographie

Bajčić, Martina / Golenko, Dejana (2024) Applying Large Language Models in Legal Translation: The State-of-the Art, in: *International Journal of Language and Law*, 13, pp. 171-196.

Barbin, Franck (2020). La traduction automatique neuronale, un nouveau tournant ?, in : *Palimpseste. Sciences, humanités, sociétés*, 4, pp. 51-53.

Biel, Łucja (2022). Translating Legal Texts, in: *The Cambridge Handbook of Translation* [ed. Kirsten Malmkjær], Cambridge: Cambridge University Press, pp. 379-400.

Biel, Łucja / Sosoni, Vilelmini (2017). The Translation of Economics and the Economics of Translation, in: *Perspectives*, 25, 3, pp. 351–361.

Boumparis, Dimitris / Giannoutsos, Christos (2023) Quantitative and Qualitative Evaluation of Human and Machine-Translated EU Economic Texts in the English-Greek Language Pair, in: *Proceedings of the New Trends in Translation and Technology Conference - NeTTT 2022. 4-6 July 2022* [ed. Sheila Castilho / Rocío Caro Quintana / Maria Stasimioti / Vilelmini Sosoni], Shouma: Incoma Ltd., pp. 248-253.

Champsaur, Caroline (2013). La traduction automatique : un outil pour les traducteurs ?, in: *The Journal of Specialised Translation*, 19, pp. 19-28.

Deng, Li / Yang Liu (2018). A Joint Introduction to Natural Language Processing and to Deep Learning, in: *Deep Learning in Natural Language Processing* [éds. Li Deng / Yang Liu], Springer : Berlin, pp. 1-22.

Ekin, Sabit (2023). Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices, in: *TechRxiv, https://www.techrxiv.org/doi/full/10.36227/techrxiv.22683919.v2*

European Language Industry Survey (2025). https://elis-survey.org/wp-content/uploads/2025/03/ELIS-2025_Report.pdf

Filipović Petrović, Ivana / Otal, Miguel López / Beliga, Slobodan (2024) Croatian Idioms Integration: Enhancing the LIdioms Multilingual Linked Idioms Dataset, in: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, pp. 4106-4112.

Gao, Yuan / Wang, Ruili / Hou, Feng (2023). How to Design Translation Prompts for ChatGPT: An Empirical Study, in: *arXiv*, https://arxiv.org/abs/2304.02182

Jiang, Zhaokun / Zhang, Ziyin (2024). Can ChatGPT rival neural machine trans-

lation? A comparative study, in: arXiv:2401.05176v1, pp. 1-20. https://arxiv.org/html/2401.05176v1

Jiao, Wenxiang / Wang, Wenxuan / Huang, Jen-tse / Wang, Xing / Shi, Shuming / Tu, Zhaopeng (2023). Is ChatGPT a Good Translator? Yes With GPT-4 as the Engine, https://arxiv.org/abs/2301.08745

Kamath, Uday / Keenan, Kevin / Somers, Garrett / Sorenson, Sarah (2024). *Large Language Models. Bridging Theory and* Practice, Berlin : Springer.

Lath, Ritika / Patwari, Renuka / Aylani, Amit / Hajoary, Deepak (2025). Introduction to Generative AI, in: *Generative AI. Disruptive Technologies for Innovative Applications* [éds. Danilo Pelusi, Nagasubramanian Gayathri, Pethuru Raj, Ramesh Chandran, S. Rakesh Kumar], Wiley: New Jersey, pp. 1-28.

Lee, Tong King (2023). Artificial intelligence and posthumanist translation: ChatGPT versus the translator, in: *Applied Linguistics Review*, 15, 6, pp. 1-22. https://doi.org/10.1515/applirev-2023-0122

Nakhlé, Mariam (2023). L'évaluation de la traduction automatique du caractère au document : un état de l'art, in : *Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)*, Paris, pp.143-159.

O'Brien, Sharon (2022). How to deal with errors in machine translation: Post-editing, in: *Machine translation for everyone: Empowering users in the age of artificial intelligence* [ed. Dorothy Kenny], Berlin: Language Science Press, pp. 51-79.

Omazić, Marija / Šoštarić, Blaženka (2023). New Resources and Methods in Translating Legal Texts: Machine Translation and Post-Editing of Machine-Translated Legal Texts, in : *Language(s) and Law* [ed. Ljubica Kordić], Osijek: Pravni fakultet Sveučilišta Josipa Jurja Strossmayera u Osijeku, pp. 71-84.

Oncevay, Arturo / Smileye, Charese H. / Liu, Xiaomo (2025). The Impact of Domain-Specific Terminology on Machine Translation for Finance in European Languages, in: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* [ed. Viviane Moreira / Anna Rogers / Michael White], Kerrville : ACL, pp. 2758-2775.

Papineni, Kishore / Roukos, Salim / Ward, Todd / Zhu, Wei-Jing (2002). Bleu: a Method for Automatic Evaluation of Machine Translation, in: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* [ed. Pierre Isabelle / Eugene Charniak / Dekang Lin], Philadelphia: Association for Computational Linguistics, pp. 311-318.

Pavlović, Nataša (2016). Strojno i konvencionalno prevođenje s engleskoga na hrvatski: usporedba pogrešaka, in : Jezik kao predmet proučavanja i jezik kao predmet poučavanja [ed. Diana Stolac / Anastazija Clastelić], Zagreb: Srednja Europa / Hrvatsko društvo za primijenjenu lingvistiku (HDPL), pp. 279-295.

Peng, Keqin / Ding, Liang / Zhong, Qihuang / Shen, Li / Liu, Xuebo / Zhang, Min / Ouyang, Yuanxin / Tao, Dacheng (2023). Towards Making the Most of ChatGPT for Machine Translation, in: *arXiv:2303.13780*. https://arxiv.org/abs/2303.13780

Petrak, Marta / Uremović, Mia / Pavelin Lešić, Bogdanka (2022). Fine-grained

human evaluation of NMT applied to literary text: case study of a French-to-Croatian translation, in: *JTDH Proceedings* [ed. Darja Fišer / Tomaž Erjavec], Ljubljana: Inštitut za novejšo zgodovino, pp. 141-146.

Rei, Ricardo / Stewart, Craig / Farinha, Ana C. / Lavie, Alon (2020). COMET: A Neural Framework for MT Evaluation, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [ed. Bonnie Webber / Trevor Cohn / Yulan He / Yang Liu], Association for Computational Linguistics, pp. 2685-2702.

Renato, Alejandro / Castaño, José / Ávila, Pilár / Berinsky, Hernán / Gambarte, Laura / Park, Hee / Pérez, David / Otero, Carlos / Luna, Daniel (2018). A Machine Translation Approach for Medical Terms, in: *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018)* [[ed. Alberto Cliquet Jr. / Sheldon Wiebe / Paul Anderson / Giovanni Saggio / Reyer Zwiggelaar / Hugo Gamboa / Ana Fred / Sergi Bermúdez i Badia], pp. 369-378.

Robert, Anne-Marie (2010). La post-édition : l'avenir incontournable du traducteur ?, in: *Traduire*, 222, https://doi.org/10.4000/traduire.460

Rossi, Caroline / Carré, Alice (2022). How to choose a suitable neural machine translation solution: Evaluation of MT quality, in: *Machine translation for everyone: Empowering users in the age of artificial intelligence* [ed. Dorothy Kenny], Berlin: Language Science Press, pp. 51-79.

Roumeliotis, Konstantinos I. / Tselikas, Nikolaos D. (2023). ChatGPT and Open-AI Models: A Preliminary Review, in: *Future Internet*, 15, 6, pp. 1-24. https://doi.org/10.3390/fi15060192

Snover, Matthew / Dorr, Bonnie / Schwartz, Rich / Micciulla, Linnea / Makhoul, John (2006). A Study of Translation Edit Rate with Targeted Human Annotation, in: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge (MA): Association for Machine Translation in the Americas.

Son, Jungha / Kim, Boyoung (2023). Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems, in : *Information 14*, 10,pp. 1-18. https://doi.org/10.3390/info14100574

De Sousa Cardoso, Cyril / Parise, Fanny (2023) *Guide de l'IA générative. Transformez votre quotidien professionnel à l'ère de ChatGPT, Bing, Bard, Bloom, Claude*, Paris : De Boeck Supérieur.

Štefanec, Vanja / Farkaš, Daša / Thakkar, Gaurish / Tadić, Marko (2024) Building a Large Language Model for Croatian. *Proceedings of New Trends in Translation and Technology Conference – NeTTT 2024* [ed. Constantin Orasan / Tharindu Ranasinghe / Gloria Corpas Pastor / Ruslan Mitkov / Maria Kunilovskaya / Vilelmini Sosoni / Marie Escribe], Varna: Incoma Ltd, pp. 204-209.

Šuman, Sabrina (2021) Overview of Natural Language Processing and Machine Translation Methods, in : *Zbornik Veleučilišta u Rijeci*, 19, 1, pp. 371-384.

Waldo, Jim / Boussard, Soline (2024). GPTs and Hallucination: Why do large language models hallucinate?, in: *Queue* 22, 4, pp. 1-15. https://doi.org/10.1145/3688007

Wiesmann, Eva (2019). Machine Translation in The Field Of Law: A Study Of The Translation Of Italian Legal Texts Into German, in: *Comparative Legilinguistics* 37, pp.117-153. https://doi.org/10.14746/cl.2019.37.4

Wisniewski, Guillaume (2025) *Traduction et IA : comment et pourquoi les giga-modèles multilingues vont révolutionner la traduction automatique ?*, https://hal.science/hal-05092920v1

Yan, Jianhao / Yan, Pingchuan / Chen, Yulong / Li, Jing / Zhu, Xianchao / Zhang, Yue (2024). Benchmarking GPT-4 against Human Translators: A Comprehensive Evaluation Across Languages, Domains, and Expertise Levels, in: *arXiv:2411.13775*, DOI 10.48550/arXiv.2411.13775

Zhang, Tianyi / Kishore, Varsha / Wu, Felix / Weinberger, Kilian Q. / Artzi, Yoav (2020) BERTScore: Evaluating Text Generation with BERT, in: *arXiv:1904.09675*. https://arxiv.org/abs/1904.09675

Textes traduits

a) Médecine (extraits)

Scheen, André J. / De Flines Jenny / Paquot, Nicolas (2023). Médicaments anti-obésité : des déceptions aux espoirs, in : *Revue Médicale de Liège*, 78, 3, pp. 147-152.

Organisation mondiale de la santé (2023). *Stratégie de riposte face à la résistance aux antipaludiques en Afrique* https://iris.who.int/bitstream/handle/10665/366302/9789240068391-fre.pdf?sequence=1

b) économie et finances (extraits)

L'oréal (2023) *Rapport financier semestriel au 30 juin 2023*, https://www.loreal-finance.com/system/files/2023-07/LOREAL_RFS_2023_FR_0.pdf

Espic, Aurélien (2025). *L'immobilier commercial en zone euro, chocs macroéconomiques ou idiosyncrasiques ?*, https://www.banque-france.fr/fr/publications-et-statistiques/publications/limmobilier-commercial-en-zone-euro-chocs-macroeconomiques-ou-idiosyncrasiques

c) droit (extraits)

Larralde, Jean-Manuel (2025). La réception du Code pénitentiaire au regard du droit européen, in : *Cahiers de la recherche sur les droits fondamentaux* , 22, https://doi.org/10.4000/12hpe

Code civil du Québec, Livre cinquième, Des obligations, https://www.legisquebec.gouv.qc.ca/fr/document/lc/ccq-1991

Annexe

Classification humaine des erreurs, version remaniée de Pavlović (2016)

A. orthographe	A.a. ponctuation
	A.b. lettre majuscule/minuscule
	A.c. autres fautes d'orthographe
B. lexique	B.a. choix lexical
	B.b. terme ou titre
	B.c. locution
C. morphosyntaxe	C.a. congruence
	C.b. formes verbales, temps verbal
	C.c. ordre des mots/parties de la phrase
D. précision	D.a. éléments non traduits
	D.b. omissions
	D.c. additions
	D.d. traduction erronée/surtraduction/sous-traduction
	D.e. chiffres, format, etc.
E. style	E.a. expressions non idiomatiques, maladroites ou trop littérales, inconsistance, registre, usage de la voix passive

Is Machine Replacing Humans? Evaluating Specialized Translations Made by Artificial Intelligence

In this article, we explored the quality of translations produced by a large language model (LLM), ChatGPT. It is a type of generative AI that is changing the translation industry. To analyze its performance, we considered three of the most frequently translated fields: law, medicine and economics. According to existing literature, machine translation (MT) has not yet achieved satisfactory results in the field of specialized texts, which is why we focused on these. The main objective was to compare human translations with those produced by ChatGPT, in order to see to what extent ChatGPT can truly help specialized translators.

To measure the quality of the translations, we used four automatic metrics (BERTScore, COMET, BLEU and TER) together with human evaluation, which consisted of classifying the errors made by ChatGPT into five categories (spelling, lexicon, morphosyntax, precision and style). Our results have demonstrated that the translations produced by ChatGPT did not achieve the desired quality, that the quality of texts in the legal field is lower than those in the other two fields, and that the largest number of errors can be found on the lexical level.

The results obtained are far from representative as they were collected from a relatively limited corpus, but they are nevertheless of significant value because they are based on a language combination (French-Croatian) that has been little explored so far in the field of MT. In conclusion, we can state that ChatGPT shows real potential for facilitating and accelerating specialized translators' work. However, as with any MT tool, careful

proofreading and post-editing are essential, especially when dealing with texts intended for publication. It should also be noted that ChatGPT's performance can be significantly improved with clear and well-written prompts. Future research would be needed to explore in more detail the effect of different types of prompts on the quality of specialized translations.

Keywords: machine translation (MT), artificial intelligence (AI), generative technology, large language models (LLM), ChatGPT, human and machine MT evaluation, specialized texts

Zamjenjuje li stroj čovjeka? Vrednovanje stručnih prijevoda izrađenih pomoću umjetne inteligencije

U ovom članku istražili smo kvalitetu prijevoda koje je proizveo veliki jezični model (LLM), ChatGPT. Riječ je o vrsti generativne inteligencije koja transformira sektor prevođenja. Kako bismo analizirali učinkovitost ChatGPT-a, razmotrili smo tri najčešće prevođena područja: pravo, medicinu i ekonomiju. Budući da prema postojećoj literaturi strojno prevođenje još uvijek ne postiže zadovoljavajuće rezultate u području stručnih tekstova, usredotočili smo se upravo na njih. Glavni cilj bio je usporediti ljudske prijevode s onima koje je proizveo ChatGPT kako bismo vidjeli u kojoj mjeri taj model zaista može biti podrška u radu prevoditelja stručnih tekstova. Za mjerenje kvalitete prijevoda koristili smo četiri automatske metrike (BERTScore, COMET, BLEU i TER) zajedno s ljudskom evaluacijom, koja se odnosila na klasifikaciju pogrešaka ChatGPT-a u pet kategorija (pravopis, leksik, morfosintaksa, preciznost i stil). Rezultati pokazuju da prijevodi koje je proizveo ChatGPT nisu postigli željenu kvalitetu, da je kvaliteta tekstova u pravnom području niža nego u ostalim dvama područjima te da se najveći broj pogrešaka nalazi u kategoriji leksičkih pogrešaka.

Dobiveni rezultati nisu sasvim reprezentativni jer su prikupljeni iz relativno ograničenog korpusa, ali su ipak od značajne vrijednosti jer se temelje na jezičnoj kombinaciji koja je slabo istražena u području strojnog prevođenja (francuski-hrvatski). Zaključno možemo reći da ChatGPT pokazuje velik potencijal za olakšavanje i ubrzavanje rada prevoditelja. Međutim, kao i kod svakog alata za strojno prevođenje, ključna je naknadna ljudska redaktura, posebno kada je riječ o tekstovima namijenjenima objavljivanju. Također treba napomenuti da se performanse ChatGPT-a mogu značajno poboljšati jasnim i dobro napisanim uputama (promptovima). Potrebna su dodatna istraživanja kako bi se detaljnije istražio utjecaj različitih vrsta promptova na kvalitetu stručnih prijevoda izrađenih pomoću umjetne inteligencije.

Ključne riječi: strojno prevođenje, umjetna inteligencija (UI), generativna tehnologija, veliki jezični modeli (LLMovi), ChatGPT, ljudsko i strojno vrednovanje strojnog prijevoda, stručni tekstovi