African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

Traduction automatique et performance des métriques BLEU et METEOR: Cas de la traduction économique

Automated Translation and the Performance of BLEU and METEOR: Economic Translation as a Case Study.

Auteur 1 : EL QUESSAR Mohammed.
Auteur 2 : NAJI Ikram.

Pr EL QUESSAR Mohammed, (Professeur de l'Enseignement Supérieur, PhD)
Université Abdelmalek Essaadi / École Supérieure Roi Fahd de Traduction – Maroc

NAJI Ikram, (https://orcid.org/0009-0006-6786-6062, MA)
Université Abdelmalek Essaadi / École Supérieure Roi Fahd de Traduction – Maroc

Déclaration de divulgation : L'auteur n'a pas connaissance de quelconque financement qui pourrait affecter l'objectivité de cette étude.

Conflit d'intérêts : L'auteur ne signale aucun conflit d'intérêts.

Pour citer cet article : EL QUESSAR .M & NAJI .I (2025) « Traduction automatique et performance des métriques BLEU et METEOR: Cas de la traduction économique», African Scientific Journal « Volume 03, Num 33 » Pp: 1322 – 1341.

DOI : 10.5281/zenodo.18099688

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

Résumé

Le présent article se veut une analyse de la performance des métriques BLEU et METEOR en fonction des aspects linguistiques des textes traduits. En effet, les nuances stylistiques, la terminologie et la structure des phrases influencent la qualité des scores BLEU et METEOR calculés. Cette réalité est particulièrement manifeste dans le domaine du discours économique. Ainsi, nous tentons de définir cet impact et de montrer comment la nature des textes économiques conditionne les résultats de l'évaluation par les scores BLEU et METEOR. Ces métriques, couramment employées dans le domaine de la traduction automatique (TA), suscitent des débats quant à leur efficacité et leur pertinence, notamment en raison de leurs limites dans l'évaluation qualitative et contextuelle des traductions produites. Cette étude vise donc à lever certaines ambiguïtés en offrant une vision plus précise et nuancée de ces outils d'évaluation.

Les outils de traduction automatique adoptés, notamment DeepL et Google Traduction, s'appuient sur des avancées majeures en apprentissage profond et en traduction automatique neuronale. Le choix de ces outils n'est pas fortuit. Il est ancré dans le contexte contemporain, marqué par l'intérêt croissant pour cette technologie. Le présent travail a donc pour ambition de fournir aux professionnels de la traduction, ainsi qu'aux chercheurs en traduction et en traductologie, une analyse critique des scores BLEU et METEOR.

Mots-clés : BLEU, METEOR, performance, évaluation quantitative, évaluation qualitative.

Abstract

This article analyzes the performance of the BLEU and METEOR metrics in terms of linguistic aspects of translated texts. In fact, style, terminology, and sentence structure are all factors that influence the quality of these scores. This is particularly true in the context of economic discourse. Commonly used in the field of machine translation (MT), these metrics spark debates regarding their effectiveness and relevance, particularly due to their limitations in assessing the qualitative and contextual aspects of the translations produced. Therefore, the present study attempts to describe this impact and show how the characteristics of economic texts are of great importance in BLEU and METEOR evaluations. Two machine translation tools were used in this study: DeepL and Google Translate. They operate using deep learning and neural machine translation and rely on significant advancements in deep learning and neural machine translation. The choice of these tools is not accidental. It is rooted in the contemporary context, characterized by growing interest in this technology. This paper targets translation professionals and researchers in translation and translation studies, aiming to provide a critical analysis of the BLEU and METEOR scores, widely used in the machine translation field but also subject to much debate. Hence, the study addresses some ambiguities by offering a more precise and nuanced view of these evaluation tools.

Keywords: BLEU, METEOR, performance, quantitative evaluation, qualitative evaluation,

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

I. Introduction

Les progrès fulgurants de l'intelligence artificielle ont révolutionné le domaine de la traduction, notamment grâce à l'émergence des techniques d'apprentissage profond et de la traduction automatique neuronale. Si ces avancées offrent de nouvelles perspectives pour une automatisation accrue des processus de traduction, elles soulèvent également des interrogations quant à leur capacité à rendre compte de la complexité et de la nuance des langues naturelles, particulièrement dans des domaines spécialisés tels que l'économie. Parallèlement, des méthodes d'évaluation dites quantitatives de ces traductions automatiques ont vu le jour. Il s'agit de métriques fondées sur des formules mathématiques rigoureuses, capables de refléter la qualité de la traduction fournie par la machine. Parmi les plus utilisées, on retrouve BLEU et METEOR. Ces métriques servent d'outil d'évaluation pour la traduction automatique. En attribuant des scores à la traduction produite par la machine, elles constituent une ressource précieuse pour le traducteur professionnel désireux de s'appuyer sur la machine pour effectuer ses tâches de traduction, tout en disposant d'outils fiables pour déterminer si la traduction produite est pertinente, adéquate et conforme aux attentes. Etant donné la popularité de la traduction effectuée par l'intelligence artificielle, et compte tenu de l'importance croissante de la traduction économique et financière, dans un contexte marqué par une mondialisation toujours plus prégnante, nous nous intéressons à comprendre le mode d'opération des métriques BLEU et METEOR afin de mieux évaluer leur efficacité pour des corpus relevant du domaine de l'économie. Cette démarche vise à offrir au traducteur professionnel une compréhension approfondie de ces outils et à lui permettre de développer une vision éclairée à leur sujet, afin de juger de leur utilité.

BLEU compare les n-grammes (séquences de mots) entre la traduction automatique et une référence humaine, tandis que METEOR prend en compte des correspondances plus profondes, telles que les synonymes ou les paraphrases. La question de la performance des métriques en tant que méthodes quantitatives d'évaluation de la traduction est d'autant plus cruciale pour les professionnels qui utilisent des technologies de traduction et pour les chercheurs en traduction et en traductologie, vu leur popularité accrue et leur utilisation de plus en plus répandue au sein de la communauté des traducteurs.

Cet article choisit donc de calculer ces deux scores pour des corpus de textes économiques, dans l'objectif de déterminer dans quelle mesure les scores BLEU et METEOR sont adaptés à la spécificité linguistique et terminologique du domaine de l'économie. Il s'efforce de répondre aux questionnements suivants : Comment les métriques BLEU et METEOR fonctionnent-elles ? Quelles sont leurs limites, notamment en ce qui concerne l'importance du contexte et de la sémantique dans la traduction ?

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

Nous avons opté pour l'application de la structure IMRAD dans le cadre de ce travail. Par conséquent, l'article est organisé de la manière suivante : dans un premier temps, nous nous attachons à définir ces métriques en exposant leur mode de fonctionnement, les équations mathématiques qui en constituent le fondement, ainsi que leurs limites. Ensuite, pour la partie méthodologique, un ensemble de 17 textes économiques a été soumis à DeepL et à Google Traduction afin d'obtenir des traductions automatiques servant de base pour le calcul des scores BLEU et METEOR. Une analyse critique des résultats de ce calcul a été menée par la suite, en prenant en considération des critères linguistiques et textuels, à savoir la longueur de la traduction de référence, le type du texte économique, et l'aspect sémantique et contextuel du texte source. En analysant l'efficacité de ces métriques, une évaluation complémentaire, effectuée par le traducteur/linguiste humain, s'est avérée nécessaire. Enfin, les résultats de cette étude ont permis d'élucider le lien entre les caractéristiques stylistiques et textuelles du corpus à traduire et la qualité de la traduction effectuée par la machine, ce qui a permis, par conséquent, de formuler des conclusions sur les scores BLEU et METEOR.

Le choix de la structure méthodologique IMRAD (Introduction, Méthodes, Résultats et Discussion) s'inscrit dans une logique épistémologique rigoureuse, adaptée à une recherche fondée sur une double évaluation quantitative et qualitative. Cette approche permet de structurer l'article de manière claire et systématique, répondant aux exigences d'une démarche scientifique. En adoptant un raisonnement hypothético-déductif, cette méthode facilite la présentation des problématiques, des outils et des résultats, tout en assurant une progression logique entre l'analyse des données quantitatives (issues des métriques d'évaluation) et leur interprétation quantitative (sous forme de tableaux de calcul des scores BLEU et METEOR) puis qualitative (Voir la partie Discussions). Elle est particulièrement pertinente dans le cadre de cette étude, qui évalue la performance des métriques dans la traduction automatique, en s'appuyant sur des indicateurs mesurables et des réflexions critiques. En ce sens, la structure IMRAD garantit une articulation cohérente entre les différentes étapes de l'analyse, tout en renforçant la transparence et la reproductibilité des résultats.

II. Cadre conceptuel des métriques BLEU et METEOR

1. Le fonctionnement des métriques

1.1. La métrique BLEU

Loin de se présenter comme une simple technique, la métrique BLEU se révèle un outil d'évaluation de la qualité de la traduction automatique qui séduit par sa finesse et sa sophistication. En effet, cette méthode quantitative, développée en 2002 par une équipe de chercheurs emmenée par Papineni, Roukos, Ward et Zhu, s'appuie sur une comparaison subtile entre les traductions

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

générées automatiquement et les références humaines, et présente ainsi une évaluation nuancée de la performance des modèles. Il s'agit d'une mesure quantitative semblable à d'autres métriques comme METEOR et TER, entre autres. La métrique BLEU calcule un score compris entre 0 et 1, et reflète le degré de correspondance entre la traduction automatique et les références, avec 1 indiquant une correspondance parfaite[1].

Bien que la métrique BLEU ne soit pas exempte de limites et de biais connus, notamment son incapacité à saisir certains aspects plus nuancés de la qualité de la traduction, elle demeure un outil précieux et largement adopté par la communauté de la traduction automatique. BLEU se distingue en effet par sa capacité à capturer la finesse de la traduction. Elle allie une précision fondée sur les n-grammes à une pénalité de brièveté afin d'éviter que les modèles ne privilégient la concision au détriment de la qualité sémantique. Dans le détail, la métrique calcule la précision de la traduction à partir de n-grammes. C'est un modèle de prédiction de mots basé sur les n-grammes qui prédit le prochain mot d'un texte à partir d'un nombre fixe (n-1) de mots précédents. Le « n » d'un n-gramme représente le nombre d'éléments (mots, caractères, etc.) pris en compte par le modèle. Cette métrique intègre également une pénalité de brièveté qui pénalise les traductions excessivement courtes par rapport aux références.

La métrique BLEU repose sur une formule qui combine une pénalité de brièveté exponentielle et une moyenne géométrique pondérée par les précisions de n-grammes. Elle permet ainsi de quantifier la similarité entre une traduction candidate et un ensemble de références. L'équation mathématique est formulée comme suit :

$$\mathbf{BLEU} = \mathbf{BP} \times \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right)$$ [2]

Où :

- w_n est le poids pour chaque n-gramme. Par exemple, $w_n = \frac{1}{4}$ pour les n-grammes jusqu'à 4.
- p_n est la précision modifiée des n-grammes de taille n. La précision modifiée mesure le nombre de n-grammes de la traduction qui apparaissent également dans la ou les références, mais elle limite le nombre de fois où un n-gramme peut être compté en prenant le minimum entre sa fréquence dans la traduction et sa fréquence maximale dans les références.

[1] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (n.d.). *BLEU: a Method for Automatic Evaluation of Machine Translation*.

[2] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (n.d.). *BLEU: a Method for Automatic Evaluation of Machine Translation*.

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

- BP est la pénalité de brièveté (Brevity penalty) qui ajuste le score pour les traductions trop courtes. Elle est calculée selon l'équation suivante :

$$\mathbf{BP} = \begin{cases} \mathbf{1, si\ c > r} \\ e^{(1-r/c)}, si\ c \leqslant r \end{cases}$$ [3]

La composante "c" reflète la longueur de la traduction candidate, et la composante "r" désigne la longueur de la traduction de référence la plus proche.

1.2. La métrique METEOR

METEOR, acronyme de "Metric for Evaluation of Translation with Explicit Ordering", se dresse tel un phare de précision dans l'océan tumultueux des systèmes de traduction automatique. Cette métrique de référence se distingue par sa finesse et sa nuance, et offre ainsi une alternative raffinée aux méthodes traditionnelles telles que le BLEU. Son éclat se révèle particulièrement dans l'évaluation des traductions à l'échelle de la phrase, où chaque mot respire la signification. METEOR se veut donc un score en synergie avec l'évaluation humaine de la qualité de la traduction. Ce score s'aligne avec les jugements humains en matière de fluidité et de pertinence, mais aussi quant à l'adéquation des segments de phrases choisis par la machine. L'évaluation des traductions s'effectue par l'alignement des mots entre le texte cible et une ou plusieurs traductions de référence.

Conçu à l'origine pour la langue anglaise, METEOR a été soigneusement ajusté afin d'améliorer ses modules de correspondance et d'adapter ses paramètres, ce qui lui permet désormais d'évaluer les traductions en espagnol, en français et en allemand. Cette métrique commence par le calcul du score F, qui combine la précision et le rappel de la traduction. Le rappel mesure la capacité d'un modèle à identifier tous les éléments pertinents. Un rappel élevé indique une couverture adéquate des éléments essentiels, tandis qu'un rappel faible signale que de nombreux éléments importants ont été omis. Le score F est calculé à travers la formule suivante:

$$F = \frac{10\ .\ P\ .R}{R + 9\ .P}$$

Ensuite, le score final (METEOR) intègre une pénalité de fragmentation. Il s'agit d'une mesure de l'ordre des mots, pénalisant les traductions où les correspondances sont très fragmentées. Enfin, la pénalité est calculée en fonction de la fragmentation.

Ci-après les formules correspondant à ces deux concepts:

METEOR = F . (1 – Penalty) [4]

[3] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (n.d.). *BLEU: a Method for Automatic Evaluation of Machine Translation.*

[4] Lavie, A., & Agarwal, A. (2007). *Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments*.

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

Penalty = $\gamma \cdot (\frac{Frag}{Matches})^{\beta}$ [5]

Où γ et β sont des paramètres ajustables déterminant la sévérité de la pénalité de fragmentation. En somme, pour évaluer la qualité d'une traduction générée par la machine, METEOR commence par établir un alignement précis entre les mots de la traduction et ceux de la référence. Cet alignement ne se limite pas aux correspondances exactes, mais intègre également les synonymes, les formes de base des mots (lemmes) et les correspondances partielles. Pour optimiser cet alignement, METEOR maximise un score combinant plusieurs critères, dont la précision (proportion de mots correctement alignés dans la traduction proposée) et le rappel (proportion de mots de la référence présents dans la traduction). Une fois l'alignement effectué, METEOR calcule un score global en pondérant différents éléments. La précision et le rappel constituent les fondamentaux de ce calcul, comme le montrent les équations mathématiques susvisées. Cependant, METEOR va plus loin en attribuant des poids spécifiques à chaque type de correspondance (exacte, par synonymie, etc.) afin de refléter leur importance relative. Enfin, la métrique pénalise les traductions qui fragmentent excessivement les mots de la référence, privilégiant ainsi des traductions plus fluides et cohérentes.

2. Limites des métriques BLEU et METEOR

2.1. Limites de la métrique BLEU

Le score BLEU revêt une importance majeure dans l'évaluation de la traduction automatique. Il est d'ailleurs très populaire et largement utilisé par les professionnels et les chercheurs en traduction. Toutefois, la métrique n'atteint ni le niveau de méticulosité ni celui d'efficacité de l'évaluation humaine. L'utilisation du score BLEU présente des défis significatifs pour le traducteur, pouvant aboutir à une évaluation imprécise des traductions générées par les systèmes automatiques. En effet, BLEU repose largement sur la forme superficielle des mots, négligeant l'aspect sémantique des phrases. Cette limitation peut entraîner des évaluations inappropriées lorsque des erreurs mineures de forme surviennent, telles que l'omission d'une négation, ce qui transforme radicalement le sens initial. Il est bien établi que deux phrases peuvent présenter des structures lexicales distinctes tout en véhiculant une signification identique. Dans ce contexte, BLEU risque de pénaliser injustement les traductions qui, bien que sémantiquement correctes, diffèrent lexicalement. La question de la négation illustre particulièrement la faiblesse majeure de BLEU, qui démontre une sensibilité insuffisante aux éléments critiques du sens, notamment aux négations et aux nuances contextuelles. Par exemple, si l'on considère la traduction de référence

[5] Lavie, A., & Agarwal, A. (2007). *Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments*.

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

"The results did not confirm the hypothesis" par rapport à une traduction automatique erronée "The results confirmed the hypothesis", il est probable que BLEU attribue un score plus élevé à cette dernière en raison de l'omission du mot "not", essentiel à la compréhension de la phrase. Par ailleurs, la métrique BLEU ne prend pas en compte les structures intermédiaires, telles que les phrases nominales ou verbales, ce qui signifie qu'elle ignore des informations contextuelles cruciales susceptibles d'influencer la qualité de la traduction. Cette lacune peut conduire à des évaluations incomplètes, car la qualité d'une traduction repose également sur les relations syntaxiques et le contexte global, et non seulement sur l'analyse des mots individuellement. Nous illustrons, dans le tableau suivant, comment BLEU peut pénaliser une traduction correcte uniquement pour la raison de sa structure syntaxique différente, en dépit de sa transmission pertinente et son maintien du sens de la phrase source.

Tableau 1 : Exemple d'une limite de la métrique BLEU

Traduction de référence	Traduction automatique "A"	Traduction automatique "B"	Commentaire
The quick brown fox jumps over the lazy dog	The fast brown fox jumps over the lazy dog	Over the lazy dog jumps the quick brown fox	● Pour la traduction A, BLEU accorde un bon score, car les mots correspondent directement à la référence, malgré le changement de "quick" en "fast". ● Pour la traduction B, BLEU peut réduire le score en raison de l'ordre des mots, même si la structure syntaxique est correcte et que le sens est préservé. ➢ La traduction "B" respecte la structure syntaxique correcte et transmet le même sens, mais BLEU ne le reconnaît pas correctement en raison de sa dépendance à l'ordre des mots. Cela montre comment BLEU peut ignorer des relations syntaxiques importantes, ce qui affecte l'évaluation de la qualité de la traduction.

Source : Réalisé par nos soins

2.2. Limites de la métrique METEOR

METEOR, bien qu'il jouisse d'une certaine popularité dans l'évaluation des traductions automatiques, peut ne pas toujours refléter fidèlement la qualité intrinsèque d'une traduction. En effet, cette métrique repose sur le principe de la meilleure hypothèse, à savoir la meilleure traduction produite par le système. Par conséquent, si les hypothèses produites ne sont pas de

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

qualité satisfaisante, le score METEOR risque de donner une image trompeuse des performances de traduction.

À l'instar de la métrique BLEU, METEOR peut également ne pas tenir suffisamment compte du contexte dans lequel les mots sont employés. Les traductions qui parviennent à s'adapter au contexte tout en restant fidèles au texte source sont souvent privilégiées. Néanmoins, une traduction trop libre par rapport à la version de référence pourrait être évaluée défavorablement, malgré sa pertinence contextuelle. Ce système de notation peut ainsi ne pas toujours saisir la subtilité et les nuances de l'expression linguistique, aboutissant à une évaluation injuste des traductions automatiques qui, aux yeux d'un évaluateur humain, seraient jugées correctes et pertinentes. Il en résulte que la qualité de l'évaluation par METEOR est fortement tributaire de celle de la traduction de référence ; si cette dernière est erronée ou ne reflète pas des traductions de haute qualité, le score attribué risque d'être peu représentatif.

Cependant, ces limites peuvent être atténuées par l'adoption de méthodes innovantes de traduction automatique, notamment les mesures de confiance. Dans son article "*On the use of confidence measures in machine translation: evaluation, post-editing and application to speech translation*", Sylvain Raybaud, docteur et chercheur en Data Science, démontre que l'intégration de ces mesures renforce la fiabilité et l'efficacité du score METEOR. Les mesures de confiance constituent des outils permettant d'évaluer et d'améliorer la fiabilité des traductions produites par un système, en estimant la probabilité qu'une traduction soit correcte ou utile. Elles sont intégrées au processus de traduction afin de fournir aux utilisateurs des scores ou des probabilités indiquant la fiabilité des traductions, un élément particulièrement crucial dans les systèmes traitant la parole spontanée, où la complexité et la variabilité du langage peuvent engendrer des incertitudes quant à la qualité des traductions.

Les limites évoquées dans cette section peuvent, sans conteste, entraver l'utilisation et la confiance accordées à la métrique METEOR. Néanmoins, ces défis sont communs à l'ensemble des métriques automatiques de traduction. Par exemple, la dépendance de METEOR aux hypothèses de traduction est logique, car les évaluations demeurent toujours conditionnées par la qualité des données d'entrée. Dans le cas où les hypothèses sont de faible qualité, les scores élevés peuvent faussement embellir les performances du système. Cette situation peut être remédiée en combinant l'évaluation par métriques et l'évaluation humaine qui, étant plus précise, fournit une vision plus exhaustive.

En conclusion, pour assurer la fiabilité des résultats obtenus, il est impératif de conduire une analyse critique approfondie des différentes métriques d'évaluation. Une telle démarche permettrait d'identifier à la fois les atouts et les biais potentiels de ces outils et de favoriser une

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

utilisation plus rigoureuse et nuancée, tout en limitant les risques d'interprétation erronée des résultats.

III. Méthodes

Dans cet article, nous évaluons la traduction automatique de textes relevant du domaine de l'économie, d'abord du Français vers l'Arabe, puis, dans un deuxième temps, de l'Anglais vers le Français. Nous avons adopté une approche en plusieurs étapes, comprenant la collecte des données, le prétraitement et l'évaluation. Nous avons décidé de traiter la traduction depuis et vers le français afin de comparer cette langue en tant que langue source puis en tant que langue cible.

Nous avons constitué un corpus à partir de textes du quotidien marocain L'Economiste. Nous avons choisi de traiter des articles de presse de ce journal parce que ce sont des textes de spécialité qui contiennent des termes, des expressions et des tournures de phrases spécifiques au domaine de l'économie. Nous avons également traité un deuxième type de texte, notamment des corpus institutionnels relatifs au marché intérieur de l'Union Européenne, sur lesquels nous avions travaillé avec l'agence de traduction French-Polish Bridges (FPB). Étant donné le caractère confidentiel de ces textes, nous avons d'abord sollicité la permission du président de l'agence en question pour les utiliser dans le cadre de ce travail.

Nous avons trié ces articles afin de sélectionner les textes contenant des phénomènes linguistiques pertinents susceptibles de poser problème à la machine. Le prétraitement est une étape cruciale pour nettoyer et préparer les données. Cela inclut des tâches telles que la division des textes en unités plus petites, notamment des paragraphes et des phrases, de manière adaptée aux objectifs de l'étude. Ainsi, nous avons obtenu 17 textes distincts que nous avons étudiés séparément.

Nous avons soumis ces textes au logiciel DeepL pour la traduction de l'Anglais vers le Français et du Français vers l'Arabe. C'est un outil de traduction automatique qui s'appuie sur le Deep Learning. Ainsi, nous avons opté pour DeepL d'abord afin d'évaluer la technologie du Deep Learning. Le choix de cet outil est également motivé par sa grande popularité et sa notoriété dans le domaine de la traduction automatique pour les résultats impressionnants qu'il génère. Il affirme d'ailleurs être le meilleur traducteur automatique au monde.

Nous avons également utilisé un deuxième outil, Google Traduction, afin de comparer les deux modèles de traduction, notamment parce que celui-ci s'appuie sur la technologie de la traduction automatique neuronale.

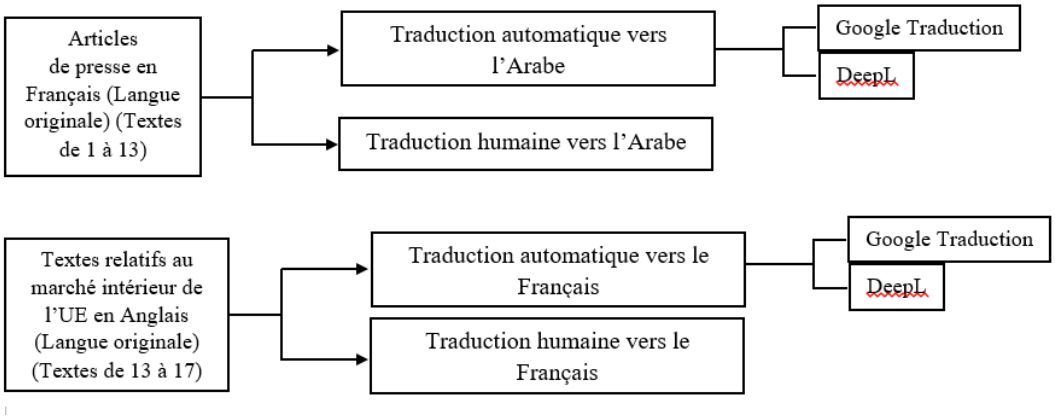
Nous avons décidé d'évaluer les traductions vers deux langues cibles différentes, notamment le Français et l'Arabe, afin de comparer le fonctionnement de la machine et les performances des scores BLEU et METEOR à l'égard de ces deux langues. La première est une langue latine utilisant les mêmes caractères que la langue source, et la deuxième, bien évidemment, une langue non

latine. Pourtant, il faut mentionner que les deux langues empruntent l'une à l'autre des mots et des expressions et s'enrichissent mutuellement.

Pour évaluer l'efficacité des traductions par la machine, nous avons utilisé deux métriques, à savoir BLEU et METEOR. Nous avons opté pour la métrique BLEU, métrique populaire pour l'évaluation des traductions, permettant une mesure pertinente de la qualité de la traduction automatique par rapport à une traduction de référence (humaine). La métrique METEOR est, elle aussi, une mesure couramment utilisée pour évaluer la qualité des résultats de la traduction automatique. Cette méthodologie nous a permis d'évaluer pertinemment les traductions effectuées par la machine, et par-dessus tout, de sortir avec des conclusions significatives pour ce qui est de l'efficacité et de la performance des métriques BLEU et METEOR.

La figure ci-après illustre la composition générale du corpus sur lequel nous avons travaillé. Elle expose la typologie textuelle que nous avons adoptée, les langues étudiées et les outils de TA utilisés pour traduire lesdits textes.

Figure : Composition générale du corpus



Source: Réalisée par nos soins

Le tableau 2 présente une comparaison quantitative du nombre de mots des traductions issues de trois sources : la traduction de référence, Google Traduction et DeepL. Cette analyse vise à quantifier l'expansion ou la contraction textuelle induite par les outils de traduction automatique. Les résultats obtenus permettent d'évaluer l'impact de ces outils sur la longueur finale des traductions et de mettre en évidence d'éventuelles différences significatives entre les deux systèmes.

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

Tableau 2 : Analyse quantitative de l'expansion/contraction textuelle lors de la traduction automatique

	Texte n°	Texte source	Traduction de référence	Traduction par Google Traduction	Traduction par DeepL
		Nombre de mots			
Du Français vers l'Arabe	**T1**	28	23	24	22
	T2	45	40	33	33
	T3	35	29	28	31
	T4	53	41	42	43
	T5	8	9	9	9
	T6	41	33	37	35
	T7	20	20	20	20
	T8	33	30	34	35
	T9	34	30	26	24
	T10	21	18	16	23
	T11	45	38	37	41
	T12	46	31	38	35
	T13	19	16	14	15
De l'Anglais vers le Français	**T14**	72	75	77	77
	T15	87	89	91	89
	T16	65	67	70	70
	T17	33	36	39	42

Source : Réalisé par nos soins

IV. Résultats

1. Évaluation automatique par scores BLEU et METEOR

Les tableaux présentés ci-après s'inscrivent dans une démarche analytique progressive. Ils visent à illustrer les résultats de notre étude comparative, articulée autour de trois axes principaux : la longueur des traductions de référence, la typologie textuelle des corpus analysés et la nature des erreurs (sémantiques et contextuelles) identifiées. Cette étude vise à quantifier l'influence de chacun de ces paramètres sur les scores BLEU et METEOR, afin d'évaluer leur impact respectif sur la qualité globale des traductions automatiques.

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

Tableau 3 : Comparaison des scores BLEU et METEOR sur la base de la longueur de la traduction de référence

Système de TA	Longueur de la traduction de référence (mots)	Score BLEU	Score METEOR
Texte 2			
Traduction de référence	40	-	-
Google Traduction	33	0,2681	0,519
DeepL	33	0,0404	0,3947
Texte 9			
Traduction de référence	30	-	-
Google Traduction	26	0,0548	0,4426
DeepL	24	0,0256	0,2919

Source : Réalisé par nos soins

Le tableau 3 présente une analyse quantitative de l'impact de la longueur des traductions sur la qualité perçue, telle qu'évaluée par les métriques BLEU et METEOR. L'hypothèse sous-jacente est que des variations significatives dans la longueur des traductions générées par Google Traduction et DeepL pourraient entraîner des variations correspondantes dans les scores obtenus, suggérant ainsi une corrélation entre la longueur et la qualité de la traduction.

Tableau 4 : Analyse comparative des performances des métriques BLEU et METEOR sur deux types de corpus : articles de presse et textes institutionnels

		Type de texte	
		Article de presse (Texte 13)	Texte institutionnel (marché intérieur de l'UE) (Texte 15)
Google Traduction	BLEU	0,0487	0,5671
	METEOR	0,4898	0,6685
DeepL	BLEU	0,0409	0,5696
	METEOR	0,3756	0,6437

Source : Réalisé par nos soins

Le tableau 4 présente les résultats de l'évaluation de la qualité de la traduction automatique pour deux types de textes : des articles de presse et des textes institutionnels. Les métriques BLEU et METEOR ont été utilisées pour évaluer la qualité des traductions produites par Google Traduction et DeepL. Cette analyse a pour objectif de déterminer si les performances de ces métriques varient selon le type de texte et de mettre en évidence les forces et les faiblesses des deux modèles de traduction.

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

Tableau 5 : Impact des erreurs sémantiques et/ou contextuelles sur les scores BLEU et METEOR

	Texte n° 3		Texte n° 7	
	Erreur contextuelle produite par DeepL	Commentaire	Erreur sémantique produite par Google Traduction	Commentaire
	بلغت القيمة السوقية 626 مليار درهم إماراتي في نهاية ديسمبر 2023	Le terme "dirham" dans la phrase source indique le dirham marocain. La machine est incapable d'identifier le contexte dans lequel est utilisée ladite phrase. La devise concernée est le dirham marocain au lieu du dirham émirati.	ونتيجة لذلك، تحول العديد من الناس عن هذا الوضع للانضمام إلى أنظمة أخرى، أو حتى العودة إلى الاقتصاد غير الرسمي.	Le terme français "statut" a été traduit par "وضع". Il s'agit d'une erreur sémantique car "statut" ici veut dire le statut juridique, qui devrait plutôt être traduit par النظام
BLEU	0,242		0,0746	
METEOR	0,5336		0,4248	

Source : Réalisé par nos soins

Le tableau 5 synthétise les résultats de l'évaluation de l'impact des erreurs sémantiques et contextuelles sur les scores BLEU et METEOR. En introduisant de manière systématique des erreurs de différents types dans des traductions de référence, nous avons cherché à évaluer la capacité des métriques à détecter et à pénaliser ces erreurs. Les résultats présentés dans ce tableau devraient permettre de mieux comprendre les limites des métriques BLEU et METEOR et d'identifier les types d'erreurs qui ont le plus d'impact sur leur évaluation.

2. Évaluation humaine

Le tableau 6 (ci-dessous) présente les résultats d'une évaluation qualitative des traductions automatiques de textes économiques, réalisées à l'aide des outils Google Traduction et DeepL. Cette étude a été menée par nous-mêmes et basée sur nos connaissances en traduction et en linguistique, conjointement à une humble expertise dans le domaine de l'économie. Les critères d'évaluation retenus portent sur la fluidité, la cohérence, la précision terminologique et la capacité à rendre compte des nuances propres au langage économique. Cette étude a pour objectif de déterminer dans quelle mesure ces outils sont capables de produire des traductions de qualité suffisante pour être utilisées dans un contexte professionnel, et d'identifier leurs forces et leurs faiblesses spécifiques en matière de traduction de textes économiques.

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

Tableau 6 : Synthèse des jugements qualitatifs sur la qualité des traductions automatiques de textes économiques par Google Traduction et DeepL

Texte original	Traduction humaine en arabe	Google Traduction	DeepL	Commentaire
De prime abord, l'AMMC explique que les indicateurs clés du marché des capitaux ont montré une tendance positive à la fin de 2023 dans un contexte économique difficile.	**مبدئيا، تشير الهيئة المغربية لسوق الرساميل أن المؤشرات الأساسية لهذا السوق أبانت عن توجه إيجابي نهاية سنة 2023 في ظل ظروف اقتصادية صعبة.**	**للوهلة الأولى، توضح الهيئة المغربية لسوق الرساميل أن المؤشرات الرئيسية لسوق رأس المال أظهرت اتجاها إيجابيا في نهاية عام 2023 في سياق اقتصادي صعب.**	**أولاً، يوضح المركز المالي الأفريقي أن مؤشرات سوق رأس المال الرئيسية أظهرت اتجاهًا إيجابيًا في نهاية عام 2023 في سياق اقتصادي صعب.**	Le logiciel de traduction DeepL a traduit "de prime abord" par "**أولا**", tandis que Google Traduction a choisi l'expression "**للوهلة الأولى**". De surplus, DeepL a traduit fautivement l'acronyme AMMC par "**المركز المالي الأفريقي**".
Les autres indices ont également progressé sur l'année, avec une augmentation de 15,4% pour le Masi 20 et de 14,3% pour le FTSE CSE Morocco 15.	**...على غرار المؤشرات الأخرى التي أحرزت تقدما خلال السنة، بارتفاع بلغت نسبته 15.4%** Masi **فيما يخص 20، و14.3% بالنسبة لمؤشر** FTSE CSE Morocco**.15**	**وتقدمت المؤشرات الأخرى أيضا خلال السنة، مع ارتفاع بنسبة 15,4%** Masi **20 لمؤشر و14,3% لمؤشر** FTSE CSE Morocco**.15**	**وارتفعت المؤشرات الأخرى أيضًا على مدار العام، مع زيادة بنسبة 15.4% لمؤشر ماسي 20 و14.3% لمؤشر** CSE **فوتسي المغرب 15.**	DeepL a transcrit en arabe les indices économiques Masi 20 et FTSE CSE Morocco 15. Résultat de la traduction : **مع زيادة بنسبة 15.4% لمؤشر ماسي 20 و14.3% لمؤشر** CSE **فوتسي المغرب** 15.
Le statut d'auto-entrepreneur peut-il encore séduire les jeunes?	**هل ما زال نظام المقاول الذاتي يلفت اهتمام الشباب؟**	**هل يمكن أن تظل حالة العمل الحر جذابة للشباب؟**	**هل لا يزال بإمكان رواد الأعمال التلقائيين جذب الشباب؟**	Les deux outils de traduction étudiés ont généré une traduction cocasse pour cette question. DeepL: **هل لا يزال بإمكان رواد الأعمال التلقائيين جذب الشباب؟** Google Traduction : **هل**

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

				يمكن أن تظل حالة العمل الحر جذابة للشباب؟
“Nous constatons que de plus en plus d’auto-entrepreneurs se retirent de ce régime, pour être éligibles au RSU afin de bénéficier d’aides directes et d’une couverture sociale», souligne le responsable d’une fiduciaire comptable.	ويشير مسؤول بشركة محاسبة: "لقد لاحظنا أن عددا متزايدا من المقاولين الذاتيين ينسحبون من هذا النظام، حتى يكونوا مؤهلين للسجل الاجتماعي الموحد، وذلك لكي يستفيدوا من الدعم المباشر والتغطية الاجتماعية".	"إننا نرى أن المزيد والمزيد من العاملين لحسابهم الخاص ينسحبون من هذا المخطط، ليكونوا مؤهلين للحصول على وحدة الدعم الاحتياطي من أجل الاستفادة من المساعدات المباشرة والتغطية الاجتماعية"، يؤكد رئيس مكتب ائتماني محاسبي.	ويشير رئيس إحدى شركات المحاسبة قائلاً: ”نحن نشهد المزيد والمزيد من أصحاب المشاريع الذاتية الذين يختارون المشاركة في هذا المخطط، لكي يكونوا مؤهلين للحصول على وحدة دعم الاستقرار والاستفادة من المساعدات المباشرة وتغطية الضمان الاجتماعي“.	La machine a traduit de manière erronée l’acronyme RSU, que ce soit pour Google Traduction ou DeepL, lesquels ont choisi pour traductions respectivement “وحدة الدعم الاحتياطي” et “وحدة دعم الاستقرار”, tandis que la traduction juste est “السجل الاجتماعي الموحد”.

Source : Réalisé par nos soins

V. Discussions

Après avoir procédé par un calcul des scores BLEU et METEOR pour 17 textes de deux différents types, nous exposons ci-après notre interprétation des résultats obtenus et nous tentons d’élaborer des conclusions quant à l’efficacité des outils de la TA, notamment en traitant des corpus dans le langage économique.

D’abord, il s’est avéré que la longueur de la traduction de référence a un impact significatif sur le score BLEU des traductions objet de cette étude. En effet, la métrique BLEU, couramment employée pour évaluer la qualité des traductions automatiques, pénalise fortement la fragmentation des traductions. Cette dernière, caractérisée par une prolifération de phrases courtes et décousues, est perçue comme un symptôme d'une traduction lacunaire, dénuée de fluidité et de cohérence. Le score BLEU, qui repose sur le comptage de séquences de mots consécutifs (n-grammes) communs entre la traduction proposée et des références humaines, voit ses scores s'éroder lorsque la traduction présente une structure fragmentée. Cette pénalisation s'explique par le fait qu'une traduction morcelée peine à saisir les nuances sémantiques et à refléter la cohésion textuelle d'un énoncé. En somme, en affaiblissant le tissage sémantique du texte, la fragmentation constitue un indicateur fiable d'une qualité traductionnelle amoindrie. Cependant, BLEU peut parfois pénaliser des traductions sémantiquement correctes mais syntaxiquement différentes des références. Il importe de ne pas perdre de vue les limites de cette métrique et d’envisager, par

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

conséquent, une combinaison de BLEU et d'autres métriques, telles que METEOR et TER, afin d'obtenir une évaluation plus pertinente de la traduction automatique.

Pour ce qui est du langage économique, les textes économiques constituent souvent un défi particulier pour les traducteurs en raison de leur richesse en terminologie, en acronymes et en expressions idiomatiques. En effet, lors du passage d'une langue à une autre, le traducteur est confronté à la nécessité de trouver un juste équilibre entre la préservation du sens original et la fluidité du texte traduit, ce qui peut parfois l'amener à opter pour des solutions qui rallongent ou, au contraire, raccourcissent le texte[6].

L'évolution sémantique des termes économiques, fortement influencée par les contextes d'utilisation, pose un défi majeur aux systèmes de traduction automatique. Ces derniers, bien que performants dans des domaines plus stables, peinent à saisir les subtilités et les nuances propres aux discours économiques. En conséquence, les traductions obtenues peuvent non seulement être inexactes, mais également entraîner des variations significatives de la longueur des énoncés, altérant ainsi le sens original. Ainsi, dans sa quête de transmettre fidèlement le message d'origine, le traducteur se trouve parfois confronté au dilemme de choisir entre deux options : d'un côté, il peut opter pour des équivalents plus explicites qui auront pour effet d'allonger le texte, et de l'autre, il peut se tourner vers des calques ou des approximations qui raccourciront le texte mais risqueront d'en altérer la précision[7]. L'ensemble de ces facteurs contribue à expliquer les difficultés rencontrées par les systèmes de TA à préserver la longueur, le style et la structure syntaxique des phrases lors du processus de traduction. Ces contraintes, inhérentes aux langues naturelles, complexifient considérablement la tâche de produire des traductions à la fois fidèles et fluides. En outre, la typologie textuelle constitue un facteur supplémentaire qui influe sur la qualité de la traduction automatique. Nous avons constaté, d'après les scores BLEU et METEOR calculés, que DeepL et Google Traduction ont rendu des traductions médiocres pour les textes de presse. Les scores ont été très faibles, particulièrement pour BLEU, et de faibles à moyens pour METEOR. Par contre, ces métriques ont donné des résultats nettement meilleurs pour les textes institutionnels relatifs au marché intérieur de l'Union européenne (UE).

Dans le tableau qui suit, nous illustrons les principales divergences entre les textes de presse et ceux institutionnels, dans une tentative de justifier la qualité médiocre de la traduction automatique pour les premiers, et la qualité moyenne à supérieure pour les seconds. Il convient, par ailleurs, de rappeler que les métriques BLEU et METEOR ne sont pas dépourvues de limites face à la

[6] Borysova, O. (2015). Some Translation Peculiarities of Economic Texts (On the Basis of Economic Texts Translation Form English into Ukrainian). *International Letters of Social and Humanistic Sciences, 64*, 162–165.

[7] Stolze, R. (2003). Vagueness in Economic Texts as a Translation Problem. *Across Languages and Cultures*, 4(2), 187-203.

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

complexité du langage et au style distingué des textes économiques. Ceci dit, ces deux outils ne sont pas entièrement transparents et ne reflètent pas la qualité de la TA de manière irréprochable.

Tableau 7: Impact des spécificités textuelles sur l'évaluation des systèmes de TA

Complexité linguistique et stylistique des textes de presse	***Nature spécifique des textes institutionnels de l'UE***
Variété lexicale et syntaxique : Le langage économique se caractérise par un lexique riche et varié, ainsi que par des constructions syntaxiques complexes, visant principalement à attirer l'attention du lecteur et à vivifier l'information communiquée. Ainsi, les outils de TA peuvent peiner à traiter cette diversité.	**Langage formel et standardisé :** Les textes institutionnels de l'UE se caractérisent par un langage formel, précis et souvent redondant. Cette uniformité facilite la tâche des modèles de traduction automatique, qui peuvent s'appuyer sur des corpus de référence de grande taille.
Ambiguïtés et nuances : Les textes de presse ne manquent pas d'ambiguïtés et d'expressions qui posent un défi pour le traducteur. Ils regorgent d'acronymes, de jeux de mots et de nuances difficiles à saisir par les modèles de langue.	**Terminologie spécialisée :** Les textes institutionnels de l'UE recourent souvent à une terminologie propre aux politiques européennes. Celle-ci est systématiquement traduite dans toutes les langues officielles de l'union.
Évolution rapide du langage : le langage journalistique évolue rapidement, avec l'apparition de nouveaux mots et d'expressions. Les modèles de traduction automatique peuvent avoir du mal à suivre cette évolution, ce qui peut nuire à la qualité des traductions.	**Structures répétitives :** Les textes institutionnels suivent souvent des structures répétitives, ce qui facilite l'identification de patterns par les modèles de traduction automatique.
Limites de BLEU et METEOR	
Sensibilité au style : Le calcul du BLEU et du METEOR est influencé par le style et la structure des phrases. Les métriques peuvent ainsi pénaliser des traductions qui, quoique correctes, ne reproduisent pas le style du texte source.	**Insuffisance pour capturer toutes les nuances :** ces métriques ne sont pas conçues pour prendre en compte l'ensemble des aspects de la qualité d'une traduction, tels que la cohérence globale du texte, la pertinence culturelle ou l'impact sur le lecteur.

Source : Réalisé par nos soins

Dans un dernier temps, nous taclons les erreurs sémantiques et contextuelles constatées dans les traductions générées par Google Traduction et DeepL. En effet, d'après le tableau 5 (voir page 13), il est clair que BLEU et METEOR ont pénalisé ces erreurs. Ceci peut être justifié par la dépendance des métriques à une correspondance exacte de mots ou de n-grammes. Elles peuvent

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

donc pénaliser des traductions qui utilisent des synonymes corrects mais différents de ceux de la référence, ou qui exploitent des sens différents d'un même mot.

Il n'en demeure pas moins que le sens d'un texte peut parfois reposer sur des informations implicites ou des inférences, ce qui est souvent le cas dans les textes économiques. Ces nuances sont difficiles à mesurer à l'aide de métriques statistiques, qui se concentrent principalement sur la correspondance explicite entre les mots. L'aspect statistique de ces outils implique également que les corpus sont évalués phrase par phrase, sans tenir compte de la cohérence globale du texte. Elles omettent également de tenir compte du contexte culturel et pragmatique dans lequel s'inscrit le texte. Une traduction linguistiquement correcte peut être inappropriée dans un contexte donné. C'est notamment le cas de l'erreur contextuelle mentionnée dans le tableau 4 (Voir page 12). Cela étant dit, il est nécessaire de compléter les métriques BLEU et METEOR par d'autres métriques linguistiques plus avancées ou par des modèles d'évaluation neuronaux. Nous avons choisi de procéder à une évaluation humaine afin de combler les lacunes de l'évaluation quantitative par BLEU et METEOR (Voir page 12).

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

VI. Conclusions

Dans cette étude, nous nous sommes intéressées à mesurer la qualité de la traduction fournie par les outils de traduction automatique (TA), notamment Google Traduction et DeepL. Nous avons choisi les textes économiques comme cas d'étude. Notre objectif principal était de déterminer l'efficacité des métriques BLEU et METEOR dans l'évaluation des traductions automatiques. En calculant les scores BLEU et METEOR pour la traduction de textes économiques, nous avons constaté que, bien que performants et utiles dans maints contextes, ils sont limités par les nuances et la subtilité du langage économique. Les aspects linguistiques propres aux corpus économiques, notamment dans le domaine de la presse et des textes institutionnels, constituent une pierre d'achoppement pour la performance de ces métriques. Ces dernières sont limitées dans leur capacité à cerner et à détecter le sens, le contexte, le style et les tournures des phrases économiques. Ainsi, elles peuvent soit pénaliser fautivement des différences de style, des synonymes ou des expressions figurées, soit, à l'inverse, avantager et valoriser des erreurs graves pouvant aller jusqu'au contre-sens.

VII. Remerciements

Je tiens à remercier Monsieur Frédéric Gabuldani-Schneider, président de French-Polish Bridges, ainsi que tous les gestionnaires de projets avec lesquels j'ai collaboré. Ils ont tous été serviables et ont joué un rôle important dans l'élaboration du corpus sur lequel je me suis appuyée pour cette étude.

African Scientific Journal
ISSN : 2658-9311
Vol : 03, Num 33 Décembre 2025

VIII. Références bibliographiques

- ASLAN, E. (2021). La Place de la Traduction Automatique dans l'Enseignement de la Traduction. *HUMANITAS - Uluslararası Sosyal Bilimler Dergisi*, *9*(18), 16–32. https://doi.org/10.20304/humanitas.944629
- *Chapitre 3_le TAL face aux données textuelles volumineuses et potentiellement dégradées*. (2015).
- Chaudiron, S. (2005). Terminologie, ingénierie linguistique et gestion de l'information. *Langages*, *157*(1), 25–35. https://doi.org/10.3917/lang.157.0025
- Fiorini, S. (2022). L'intelligence artificielle au défi du multilinguisme : usages et perspectives de la traduction automatique neuronale dans la communication scientifique. *I2D - Information, Données & Documents*, *n° 1*(1), 73–76. https://doi.org/10.3917/i2d.221.0073
- Langevin, C. (2022). Les technologies de l'intelligence artificielle au service des médias et des éditeurs de contenus : traitement du langage naturel (TAL). *I2D - Information, Données & Documents*, *n° 1*(1), 30–37. https://doi.org/10.3917/i2d.221.0030
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*.
- Poibeau, T. (2014). LE TRAITEMENT AUTOMATIQUE DES LANGUES POUR LES SCIENCES SOCIALES: Quelques éléments de réflexion à partir d'expériences récentes. *Reseaux*, *188*(6), 25–51. https://doi.org/10.3917/res.188.0025
- Raybaud, S. (2012). *On the use of confidence measures in machine translation : evaluation, post edition and application to speech translation*. https://www.researchgate.net/publication/297303204
- Sennrich, R., Haddow, B., & Birch, A. (2015). *Improving Neural Machine Translation Models with Monolingual Data*. http://www.statmt.org/wmt15/
- Traduction automatique et usage linguistique : Une analyse de traductions anglais-français réunies en corpus. (2018). *Meta (Canada)*, *63*(3), 786–806. https://doi.org/10.7202/1060173ar
- Yan, J., Meng, F., & Zhou, J. (2020). *Multi-Unit Transformers for Neural Machine Translation*. Association for Computational Linguistics. https://github.com/Ellio