Étude préliminaire des problèmes liés à la variation terminologique dans les sorties de traduction automatique : De l'annotation de la variation à l'annotation d'erreurs

José Cornejo Cárcamo

Étude préliminaire des problèmes liés à la variation terminologique dans les sorties de traduction automatique

De l'annotation de la variation à l'annotation d'erreurs

José CORNEJO CÁRCAMO[1]
Sous la direction de Natalie KÜBLER[1], Alexandra MESTIVIER[1]
[1]ALTAE – Université Paris Cité
Séminaire LSCT – 16 février 2026

Introduction

[Diapositive 1]

Bonjour à toutes et à tous. Je vous remercie de m'accueillir dans le cadre de ce séminaire LSCT. Je m'appelle José Cornejo Cárcamo, je suis doctorant au sein du laboratoire ALTAE à l'Université Paris Cité, et je travaille sous la direction de Natalie Kübler et Alexandra Mestivier. La communication que je vous propose aujourd'hui s'intitule *« Étude préliminaire des problèmes liés à la variation terminologique dans les sorties de traduction automatique : de l'annotation de la variation à l'annotation d'erreurs »*. Cette présentation s'inscrit dans le cadre du projet ANR MaTOS (*Machine Translation for Open Science*), qui vise à développer de nouvelles méthodes pour la traduction automatique intégrale de documents scientifiques.

[Diapositive 2]

Ma présentation se structure en sept parties. Je commencerai par un cadrage théorique sur la variation terminologique et les défis qu'elle pose aux systèmes de traduction automatique. Je présenterai ensuite la typologie de la variation dénominative que nous avons développée, puis le corpus annoté d'articles scientifiques en sciences de la Terre et des Planètes. J'exposerai la méthodologie d'annotation et les résultats obtenus. Dans un second temps, je présenterai la typologie que nous avons élaborée pour l'annotation des erreurs terminologiques dans les sorties de traduction automatique, ainsi que les premières erreurs repérées. Je conclurai par les perspectives de cette recherche.

I. La variation terminologique : un défi pour les systèmes de traduction automatique

1.1. La variation en terminologie

[Diapositive 3]

Les linguistes s'intéressent à la variation terminologique depuis plus de vingt ans. Des travaux fondateurs, notamment ceux de Cabré (1999), Collet (2004) et Desmet (2006), ont contribué à démontrer la présence et l'usage des variantes dénominatives dans le discours spécialisé. Plus récemment, Fernández-Silva (2019), Freixa (2022) et Biel (2023) ont précisé que la variation constitue un phénomène naturel et inhérent aux langues de spécialité. Ce phénomène n'est donc pas un accident ni une anomalie : il fait partie intégrante du fonctionnement des terminologies dans leur contexte d'usage.

[Diapositive 4]

Du point de vue théorique, la variation terminologique peut être définie comme un phénomène linguistique provoquant toute modification portant sur la forme ou le fond d'un terme (Fernández-Silva, 2019 ; Freixa & Fernández-Silva, 2017). Elle est souvent divisée en deux grands types, comme le rappellent Carrasco et León Araúz (2023) : la variation dénominative, d'une part, et la variation conceptuelle, d'autre part. Il convient de souligner, à la suite de León Araúz (2017), que ces deux types de variation peuvent se produire simultanément ou découler l'un de l'autre, ce qui rend le phénomène d'autant plus complexe à appréhender.

1.2. La variation dénominative

[Diapositive 5]

La variation dénominative correspond à la présence de plusieurs termes différents pour désigner une même notion ou un même concept dans un domaine spécialisé, comme l'a défini Sager (1990). Elle est souvent liée à la synonymie et considérée comme un signe de néonymie, c'est-à-dire de création terminologique (Gledhill & Pecman, 2018). De nombreuses études ont été menées sur ce phénomène, parmi lesquelles celles de Freixa (2002), Suárez de la Torre (2004), Seghezzi (2011), Pecman (2012), Daille (2017), Alves da Costa et Fernández-Silva (2018), Biel (2023), Carrasco et León Araúz (2023) et Candel Mora (2024). Ces travaux ont donné lieu à des typologies différentes, qu'il est possible de fusionner pour obtenir une vision plus complète du phénomène.

[Diapositive 6]

Le plan dénominatif met en évidence plusieurs dimensions de la variation. Premièrement, il permet d'observer les changements et la flexibilité entre les différentes dénominations, qu'il s'agisse de changements de forme relevant de la morphosyntaxe ou de changements lexicaux impliquant la synonymie des constituants. Deuxièmement, il invite à interroger les possibles motivations à l'origine de la variation : des motivations cognitives, liées à la multidimensionnalité des concepts ; des motivations fonctionnelles, liées à l'adaptation au public cible ; ou encore des motivations dialectales, liées à l'appartenance à une communauté disciplinaire. Troisièmement, il permet d'identifier les fonctions associées à l'usage des variantes, qu'elles soient cognitives (refléter différentes perspectives sur un même concept) ou discursives (éviter la répétition dans le texte).

1.3. Un défi pour les systèmes de traduction automatique

[Diapositive 7]

La variation terminologique pose un problème important aux systèmes de traduction automatique. Comme l'ont souligné Bawden et al. (2025), Bénard (2023) et Cabezas García et León Araúz (2023), plusieurs facteurs contribuent à cette difficulté. D'abord, les ressources et les métriques pour identifier les faiblesses des systèmes face à la variation sont encore insuffisantes. Ensuite, les termes font souvent partie de structures complexes, notamment des groupes nominaux complexes (*complex noun phrases*), dont la traduction représente un défi supplémentaire. Par ailleurs, les nouveaux termes du domaine manquent de stabilité, ce qui complique leur traitement par les systèmes. Enfin, la variation intra-textuelle, c'est-à-dire

l'usage de plusieurs dénominations pour un même concept au sein d'un même texte, n'est généralement pas modélisée par les systèmes de traduction automatique.

II. Typologie de la variation dénominative

[Diapositive 8]

La typologie que je vais maintenant vous présenter a été développée dans le cadre du projet ANR MaTOS. Ce projet vise, entre autres objectifs, à développer de nouvelles méthodes pour la traduction automatique intégrale de documents scientifiques tout en traitant la variation terminologique. Il s'agit là d'un enjeu important pour les langues de spécialité, dans la mesure où les systèmes de traduction automatique actuels ne prennent généralement pas en compte ce phénomène. Le projet vise également à développer des métriques automatiques pour évaluer la qualité des traductions produites.

[Diapositive 9]

Notre typologie de la variation dénominative comprend cinq grandes catégories. La première est la variation graphique (VG), qui recouvre plusieurs sous-types : l'alternance entre un terme complexe et un acronyme, un sigle ou une abréviation (par exemple, acoustic Doppler current profiler qui alterne avec ADCP) ; l'alternance entre un terme et une forme artificielle telle qu'un symbole ou une formule chimique (amorphous silicon alternant avec a-Si) ; les changements orthographiques simples portant sur le tiret, l'espace ou la casse (aeolian bedform face à eolian bedform) ; les changements orthographiques multiples (organic carbon burial face à organic-C burial) ; et enfin l'abréviation de certains constituants d'un terme complexe (apparent polar wander path réduit à APW path).

La deuxième catégorie est la variation morphosyntaxique (VMS), qui comprend le changement de l'ordre des constituants (*thrust and fold system* face à *fold-and-thrust system*), le maintien de la structure avec ou sans article, les changements de flexion (*buckle* face à *buckling*), les changements morphologiques (*accretion model* face à *accretionary model*) et les changements de structure syntaxique (*acoustic-gravity wave* face à *gravity and acoustic wave*).

[Diapositive 10]

La troisième catégorie est la variation par réduction (VR), qui comprend la réduction de la base (*forearc zone* réduit à *forearc*), la réduction de l'extension (*accretionary prism* réduit à *prism*) et d'autres types de réductions (*acoustic ballast release system* réduit à *acoustic release*).

La quatrième catégorie est la variation par expansion (VE). Elle inclut l'ajout d'un trait sémantique ou attribut, pertinent mais superflu, comme dans *slow slip* face à *slow fault slip*, ou *abrasion* face à *mechanical abrasion* ; la substitution d'un morphème ou d'un lexème formant une unité complexe (*phyllosilicate* face à *sheet silicate*) ; et l'insertion lexicale, qu'elle soit nominale, verbale ou paraphrastique (*biopolymers* face à *bio-based polymers*).

Enfin, la cinquième catégorie est la variation lexicale (VL), qui peut concerner une unité simple (*anhedral* face à *xenomorphic*), une unité complexe avec changement de la base (*slow slip event* face à *slow slip episode*), un changement de l'extension (*abyssal plain* face à *deep-sea plain*), ou encore un changement portant à la fois sur la base et sur l'extension (*slow slip* face à *fault creep*).

III. Corpus annoté d'articles scientifiques

[Diapositive 11]

Le corpus sur lequel nous travaillons est composé de dix articles de recherche rédigés en anglais, relevant du domaine des sciences de la Terre et des Planètes. Il totalise environ 117 750 mots. Les articles proviennent de différentes revues spécialisées, telles que Earth-Science Reviews, Earth and Planetary Science Letters, Precambrian Research, Tectonophysics, Geochemistry, Geophysics, Geosystems, Sedimentary Geology et le Journal of Structural Geology. Ce corpus est utilisé pour deux tâches distinctes : d'une part, l'annotation de la variation terminologique et, d'autre part, l'annotation des erreurs dans les sorties de traduction automatique. Les articles ont été choisis à partir du travail déjà effectué pour la base ARTES, un outil d'aide à la rédaction et à la traduction en langues de spécialité (Kübler & Pecman, 2012 ; Pecman, 2021).

IV. Méthodologie et résultats

4.1. Identification des termes de base

[Diapositive 12]

La méthodologie que nous avons adoptée pour l'annotation de la variation repose, dans un premier temps, sur une lecture intégrale de chaque article, ce qui constitue une approche essentiellement manuelle. L'identification des termes de base s'effectue d'abord à partir du titre, du résumé et des mots-clés de l'article, conformément à la démarche préconisée par Fernández-Silva (2016, 2019). Par la suite, d'autres termes de base et variantes sont identifiés au fil de la lecture. Le terme de base est défini comme la première occurrence lexicale d'un concept dans le texte, tandis que les variantes correspondent aux occurrences ultérieures faisant référence au même concept sous une forme différente.

4.2. Recherche de marqueurs discursifs

[Diapositive 13]

[Diapositive 14]

En complément de cette approche manuelle, nous recourons à une approche semi-automatique consistant à rechercher des marqueurs discursifs dans le corpus à l'aide de Sketch Engine. Ces marqueurs, identifiés notamment à partir des travaux de Delavigne (2020) et de la thèse de Suárez de la Torre (2004), établissent un lien explicite entre un terme de base et une variante, voire entre deux variantes, selon leur ordre d'apparition dans le texte. Parmi les marqueurs recherchés figurent des expressions telles que *called*, *known as*, *named*, *or*, les parenthèses, *referred to as*, *termed* ou encore *viewed as*. Par exemple, dans la phrase « *slow slip, also referred to as fault creep, can be many orders of magnitude slower* », le marqueur *also referred to as* relie explicitement les termes *slow slip* et *fault creep*, confirmant leur relation de variation. De même, l'usage des parenthèses dans « *slower (aseismic) slip* » permet d'identifier un lien entre *aseismic slip* et *slow slip*.

4.3. Recherche par constituants

[Diapositive 15]

Une troisième méthode consiste en une recherche contextuelle pour chaque constituant d'un terme ou d'une variante, dans le but de repérer d'autres variantes possibles. Cette approche s'appuie sur les travaux de Hamon et Nazarenko (2001) ainsi que de Fernández-Silva (2011, 2016, 2019). Le principe est le suivant : un lien de synonymie pourrait exister entre deux termes complexes s'ils partagent la même construction syntaxique et des constituants synonymes. Ainsi, en recherchant systématiquement les contextes d'apparition de chaque constituant, il est possible de mettre au jour des relations de variation qui ne seraient pas immédiatement repérables par la seule lecture linéaire du texte.

4.4. Validation des termes et des variantes

[Diapositive 16]

La validation des termes et des variantes identifiés constitue une étape essentielle de notre méthodologie. Elle repose sur plusieurs types de ressources. Nous consultons des bases de données terminologiques, en particulier Termium Plus et la base ARTES (Kübler & Pecman, 2012 ; Pecman, 2021). Nous recourons également à Google Scholar pour vérifier l'usage des termes dans la littérature scientifique. Enfin, nous exploitons un corpus comparable interne à l'Université Paris Cité, constitué de textes en géosciences et totalisant environ 71 millions de mots en anglais et 192 millions de mots en français (Kübler et al., 2024). L'objectif de cette validation est de confirmer que les formes identifiées correspondent bien à un usage attesté par des spécialistes du domaine.

4.5. Annotation sur Human Signal

[Diapositive 17]

L'annotation proprement dite est réalisée sur l'interface Human Signal (Tkachenko et al., 2020). Cette interface permet d'établir visuellement les liens entre les termes de base et les variantes. Concrètement, l'annotateur identifie et surligne chaque terme de base en lui attribuant l'étiquette « Terme vedette », puis identifie chaque variante en lui attribuant l'étiquette « Variante » accompagnée du sous-type de variation correspondant (par exemple, VG changement base extension, VE trait sémantique, VMS changement structure, etc.). Des relations sont ensuite tracées entre chaque terme de base et ses variantes, ce qui permet de visualiser l'ensemble du réseau de variation au sein de chaque texte.

4.6. Distribution des données de la variation

[Diapositive 18]

Les résultats de l'annotation font apparaître 464 paires distinctes terme de base/variante à travers l'ensemble du corpus. Il est important de préciser que ce chiffre reflète la diversité des relations de variation, et non la fréquence d'usage de chaque type de variante dans les textes. En termes de distribution par catégorie principale, la variation graphique (VG) arrive en tête avec 157 paires (33,8 %), suivie des changements multiples, notés CM dans nos graphiques (94 paires, soit 20,3 %), de la variation lexicale (VL, 79 paires, soit 17 %), de la variation par réduction (VR, 76 paires, soit 16,4 %), de la variation par expansion (VE, 38 paires, soit 8,2 %) et enfin de la variation morphosyntaxique (VMS, 20 paires, soit 4,3 %).

[Diapositive 19]

Si l'on considère non plus les paires distinctes, mais les occurrences annotées, le tableau est différent. Au total, 1 924 occurrences de variantes ont été relevées dans le corpus. La variation graphique domine nettement avec 1 041 occurrences (54,1 %), ce qui s'explique par la très haute fréquence de l'alternance entre les termes complexes et leurs acronymes ou sigles dans les articles scientifiques. La variation par réduction représente 316 occurrences (16,4 %), les changements multiples 228 occurrences (11,9 %), la variation lexicale 191 occurrences (9,9 %), la variation par expansion 89 occurrences (4,6 %) et la variation morphosyntaxique 59 occurrences (3,1 %). Les occurrences montrent ainsi la véritable fréquence pour chaque type de variation dans le corpus.

[Diapositive 20]

Un résultat important est que les variantes sont globalement plus fréquentes que les termes de base dans le corpus : on relève 1 924 occurrences de variantes contre 1 492 occurrences de termes de base. Ce constat confirme que la variation terminologique n'est pas un phénomène marginal dans la rédaction scientifique. Les auteurs utilisent activement des formes alternatives pour désigner les concepts de leur domaine, ce qui constitue un défi important pour les systèmes de traduction automatique, qui doivent être capables de reconnaître et de traiter cette diversité dénominative.

[Diapositive 21]

La répartition par texte révèle des profils contrastés. Certains textes, comme les textes 3, 4, 5, 6, 9 et 10, présentent un usage très marqué des variantes, tandis que dans d'autres, comme les textes 1, 2, 7 et 8, les termes de base restent plus fréquents. Ces écarts reflètent probablement des différences de style rédactionnel : certains auteurs privilégient la reprise des formes de base, tandis que d'autres diversifient davantage leurs formulations. Cette observation a des implications directes pour l'évaluation de la traduction automatique, car elle montre que le degré de variation varie considérablement d'un texte à l'autre.

4.7. Exemples de termes et variantes identifiés

[Diapositive 22]

Parmi les paires terme de base/variante les plus fréquentes en termes d'occurrences, on trouve des cas de variation graphique tels que *Carboneras Fault Zone* alternant avec *CFZ* (146 occurrences), *vertical gravity gradient* avec *Tzz* (105 occurrences), ou *Uunimäki gabbro* avec *UGB* (75 occurrences). Les variations graphiques qui prédominent sont donc liées à une alternance entre des termes complexes et des sigles ou des formes artificielles. Les variantes par réduction concernent notamment la réduction de l'extension, impliquant le passage vers une forme hyperonymique, comme dans *forearc basin* réduit à *basin* (40 occurrences). On observe également des cas de variation morphosyntaxique, de variation par expansion et de variation lexicale parmi les paires les plus fréquentes.

[Diapositive 23]

Les données mettent également en lumière le nombre de variantes associées à chaque terme de base. Certains termes présentent un réseau de variation particulièrement riche. Ainsi, le terme

slow slip event possède 14 variantes distinctes, *metamorphic basement blocks* en possède 9, *subduction interface* et *orogenic tectonic stress* en possèdent chacun 8. Cette diversité de formes pour un même concept illustre la richesse et la complexité du phénomène de variation dénominative dans le discours scientifique.

V. Typologie pour l'annotation d'erreurs terminologiques

[Diapositive 25]

Je passe à présent à la seconde partie de ma présentation, consacrée à l'annotation des erreurs terminologiques dans les sorties de traduction automatique. La méthodologie adoptée comprend plusieurs étapes : la préparation d'une typologie d'erreurs focalisée sur les problèmes terminologiques ; la traduction automatique des dix articles scientifiques à l'aide du système de TA développé dans le cadre du projet MaTOS (Peng et al., 2025 ; Bénard et al., 2023) ; l'annotation sur Human Signal des termes de base et des variantes dans les segments traduits ; la vérification des propositions du système sur corpus, bases de données terminologiques et Google Scholar ; et enfin le traitement des données extraites au format JSON à l'aide d'un code développé sur R Studio.

[Diapositive 26]

La typologie d'erreurs terminologiques que nous avons élaborée s'appuie sur les travaux de Bénard (2024), Bénard, Kübler, Mestivier, Minder et Zhu (2024), Cabezas García et León Araúz (2023), Kübler, Mestivier et Pecman (2022) et Kübler (2008). Elle est structurée en trois niveaux afin que l'annotation soit aussi fine que possible. Un score de gravité est attribué à chaque type d'erreur, permettant de pondérer les erreurs en fonction de leur impact sur la qualité de la traduction.

[Diapositive 27]

La typologie comprend huit catégories principales, que je vais détailler : A. Sélection incorrecte d'équivalent terminologique ; B. (In)cohérence terminologique dans un texte ; C. Erreurs liées aux constituants des termes complexes ou des groupes nominaux complexes ; D. Erreurs liées aux relations sémantiques ; E. Erreurs de transposition ; F. Erreurs grammaticales ou phraséologiques ; G. Altération du contenu ; H. Expression en langue cible.

5.1. Sélection incorrecte d'équivalent terminologique (catégorie A)

[Diapositive 28]

[Diapositive 29]

La première catégorie, la sélection incorrecte d'équivalent terminologique, comprend sept sous-types. Le sous-type A1 correspond à un terme traduit par un autre terme du domaine, inexact : par exemple, *tremor* traduit par « tremblement de terre » au lieu de « trémor ». Le sous-type A2 concerne un terme traduit par une paraphrase de langue générale, comme *lithospheric mantle* rendu par « couche profonde sous la croûte terrestre » au lieu de « manteau lithosphérique ». Le sous-type A3 porte sur les créations littérales du système, comme *rate-and-state friction* traduit par « taux et état de frottement » au lieu de « friction à variable d'état ». Le sous-type A4 concerne la non-traduction d'un terme qui devrait être traduit, comme

megathrusts laissé tel quel au lieu d'être rendu par « mégachevauchements ». Le sous-type A5 traite du cas inverse : la traduction d'un terme qui ne devrait pas l'être, comme MORB traduit par « BDRM ». Le sous-type A6 porte sur le choix lexical incorrect d'un constituant, comme la confusion entre « réflexion » et « réfraction » dans *active source refraction seismology*. Enfin, le sous-type A7 correspond à un terme traduit par un mot de langue générale, comme *repeaters* traduit par « répétiteurs » au lieu de « séismes répétitifs ».

5.2. (In)cohérence terminologique (catégorie B)

[Diapositive 30]

La deuxième catégorie concerne l'(in)cohérence terminologique au sein d'un même texte traduit. Le sous-type B1 correspond au cas où un même terme de base reçoit plusieurs traductions différentes, comme *earthquake rupture* traduit tantôt par « rupture sismique », tantôt par « ruptures de séismes ». Le sous-type B2 concerne les traductions multiples d'une même variante, comme *fault creep* traduit alternativement par « fluage asismique », « fluage de la faille » ou « fluage de faille ». Le sous-type B3 survient lorsque le terme de base et une variante sont traduits de la même manière, effaçant ainsi la distinction conceptuelle que l'auteur avait établie entre les deux formes. Le sous-type B4 concerne le cas où différentes variantes sont traduites de façon identique.

5.3. Erreurs liées aux constituants (catégorie C)

[Diapositive 31]

La troisième catégorie regroupe les erreurs liées aux constituants des termes complexes ou des groupes nominaux complexes. Le sous-type C1 correspond à une identification erronée de la tête du terme, comme dans *rate and state friction* traduit par « taux à friction et état ». Le sous-type C2 concerne le rattachement d'un modifieur au mauvais constituant, comme *precursory earthquake swarms* traduit par « essaims de précurseurs d'un séisme » au lieu de « essaims de séismes précurseurs ». Le sous-type C3 porte sur l'ordre erroné des constituants, comme *slow earthquake* traduit par « lent séisme » au lieu de « séisme lent ».

5.4. Erreurs liées aux relations sémantiques (catégorie D)

[Diapositive 32]

La quatrième catégorie traite des erreurs liées aux relations sémantiques. Le sous-type D1 concerne la restitution erronée des liens sémantiques entre constituants. Par exemple, *fast and slow events* traduit par « événements à la fois rapides et lents » modifie le sens : il s'agit en réalité de deux types d'événements distincts, et non pas d'événements possédant simultanément les deux propriétés. Le sous-type D2 concerne la factorisation non détectée d'un constituant dans une coordination, comme dans *pressure-solution and dislocation creep*, où le terme *creep* s'applique aux deux éléments coordonnés, ce que le système n'a pas identifié.

5.5. Erreurs de transposition (catégorie E)

[Diapositive 33]

La cinquième catégorie porte sur les erreurs de transposition. Le sous-type E1 concerne l'absence d'explicitation nécessaire lors du passage d'une structure synthétique à une structure analytique. Par exemple, *strike-slip and subduction thrust faults* nécessite en français une explicitation du type « des failles de décrochement et de chevauchement dans les zones de subduction ». Le sous-type E2 concerne les calques erronés, comme *SSEs* (au pluriel) conservé tel quel en français au lieu d'être adapté en « SSE » (invariable).

5.6. Erreurs grammaticales ou phraséologiques (catégorie F)

[Diapositive 34]

La sixième catégorie concerne les erreurs grammaticales ou phraséologiques. Le sous-type F1 porte sur les problèmes de détermination, de nombre ou de genre, comme *fault creep* traduit par « la glissement de failles » (erreur de genre et de nombre) au lieu de « le glissement de faille ». Le sous-type F2 concerne l'absence ou l'erreur de préposition, comme « fluides de haute pression » au lieu de « fluides à haute pression ». Le sous-type F3 porte sur les collocations ou la phraséologie erronée pour le domaine. Le sous-type F4 concerne les erreurs d'orthographe, comme « glissement aseismique » au lieu de « glissement asismique ».

5.7. Altération du contenu et expression en langue cible (catégories G et H)

[Diapositive 35]

La septième catégorie traite de l'altération du contenu. Le sous-type G1 porte sur l'ajout injustifié d'une unité lexicale ou d'un syntagme, comme *fault creep* traduit par « glissement lent de la faille », où l'adjectif « lent » est absent du texte source. Le sous-type G2 concerne l'omission d'un constituant, comme *seismicity bursts* réduit à « séismes ». Le sous-type G3 porte sur les traductions inintelligibles ou les hallucinations, comme *SSEs* traduit par « événements sismiques superficiels ». Enfin, la huitième catégorie (H) concerne l'expression en langue cible, notamment l'inadaptation des poids des constituants (H1) et le style ou registre inapproprié (H2), comme *earthquake swarms* rendu par « essaims de tremblements de terre » au lieu de « essaims sismiques », forme plus adaptée au registre scientifique.

5.8. Scores de gravité

[Diapositive 36]

Chaque erreur se voit attribuer un score de gravité sur une échelle de 0 à 3, en nous appuyant sur les travaux de Minder (2024) et Bénard et al. (2024). Le score 0 (neutre) indique qu'une meilleure traduction pourrait être proposée, mais que la traduction retenue ne constitue pas réellement une erreur et n'impacte en aucun cas la compréhension. Le score 1 (mineur) signifie que l'erreur a un très léger impact sur le texte cible sans nuire à la lisibilité. Le score 2 (majeur) signale une erreur affectant significativement la compréhension, la lisibilité ou la pertinence du contenu, entraînant une perte ou un glissement de sens. Enfin, le score 3 (critique) indique que l'erreur rend le contenu totalement faux ou inexploitable, nécessitant une reformulation totale et constituant un obstacle majeur à l'utilisation du texte.

VI. Les erreurs repérées jusqu'à présent

[Diapositive 37]

La tâche d'annotation des erreurs est en cours. Un seul texte a été intégralement annoté à ce stade, ce qui nous permet néanmoins de mettre en pratique la typologie et d'observer les premiers résultats. Cette phase initiale sert également à affiner la typologie en fonction des erreurs effectivement observées et à mener des discussions au sein de l'équipe pour choisir la ou les étiquettes les plus appropriées pour chaque erreur.

[Diapositive 38]

Le premier texte annoté compte 12 646 mots, dont 609 segments annotés. Sur ces 609 segments, 360 (59 %) ne comportent aucune erreur terminologique, tandis que 249 segments (41 %) présentent au moins une erreur. La catégorie la plus représentée est la sélection incorrecte d'équivalent terminologique (catégorie A), avec 95 segments concernés (16 %), suivie de l'incohérence terminologique (catégorie B) avec 44 segments (7 %), de l'altération du contenu (catégorie G) avec 37 segments (6 %), des erreurs de transposition (catégorie E) et des erreurs liées aux constituants (catégorie C) avec 20 segments chacune (3 %), de l'expression en langue cible (catégorie H) avec 14 segments (2 %), des erreurs liées aux relations sémantiques (catégorie D) avec 10 segments (2 %), et enfin des erreurs grammaticales ou phraséologiques (catégorie F) avec 9 segments (1 %). On observe donc que la sélection d'équivalent constitue la principale source de difficultés pour le système.

[Diapositive 39]

L'analyse du taux d'erreur selon que le segment contient un terme de base ou une variante révèle un résultat éclairant. Pour les termes de base, sur 297 occurrences, 189 ont été traduites correctement (64 %) et 108 comportent une erreur (36 %). Pour les variantes, sur 311 occurrences, 170 ont été traduites correctement (55 %) et 141 comportent une erreur (45 %). Le taux d'erreur est donc sensiblement plus élevé pour les variantes (45 %) que pour les termes de base (36 %), ce qui suggère que la variation terminologique représente une difficulté accrue pour le système de traduction automatique.

[Diapositive 40]

L'examen des traductions erronées les plus fréquentes permet d'identifier les cas les plus problématiques. Les erreurs les plus récurrentes (8 occurrences chacune) concernent la variante *repeaters* traduite par « répétiteurs » (erreurs de type A6 choix lexical et A7 mot de langue générale), la variante *rate-and-state friction* traduite par « friction à taux et état » (erreurs A3 création littérale et C1 erreur de tête. On relève également la variante *slow-slip* traduite par « glissement lent » de la même manière que le terme de base (erreur B3, 6 occurrences) ou encore le sigle *SSEs* conservé tel quel en français (erreur E2 calque, 5 occurrences). Ces résultats montrent que les erreurs touchent aussi bien les termes de base que les variantes, et qu'elles combinent souvent plusieurs catégories d'erreurs.

[Diapositive 41]

L'analyse des incohérences de traduction par famille terminologique révèle 44 incohérences au total dans ce premier texte, dont 14 de type B1 (différentes traductions d'un même terme de base) et 29 de type B3 (même traduction pour le terme de base et sa variante). La famille terminologique *slow slip* est la plus touchée, avec 10 incohérences de type B3 : au moins l'une

des variantes de *slow slip* a été traduite dix fois de la même manière que le terme de base. La famille *earthquake rupture* présente également 10 incohérences, réparties entre B1 et B3 : le terme de base a reçu sept traductions différentes (B1), et au moins l'une de ses variantes a été traduite trois fois de la même manière que le terme de base (B3).

[Diapositive 42]

Pour illustrer plus finement ces phénomènes d'incohérence, nous avons préparé des schémas de correspondance entre les termes sources et leurs traductions. Le premier schéma porte sur la famille terminologique de *slow slip*. On y observe que le terme de base *slow slip* et ses variantes (*aseismic fault slip*, *aseismic slip*, *fault creep*, *slow fault slip*, *slow-slip*) convergent massivement vers la traduction « glissement lent » (137 occurrences correctes, mais 8 occurrences de type B3). Le système produit ainsi une neutralisation de la variation : des formes sources distinctes, porteuses de nuances sémantiques différentes, sont ramenées à un même équivalent en français. D'autres traductions correctes apparaissent plus marginalement, comme « glissement asismique » (14 occurrences correctes) ou « glissements lents » (8 occurrences correctes, 2 B3).

[Diapositive 43]

Le second schéma porte sur la famille terminologique de *earthquake rupture*. On compte 10 variations dans la sortie de traduction, avec 7 incohérences de type B1 et 3 de type B3. Le terme de base *earthquake rupture* est traduit par « rupture sismique » (3 occurrences correctes, 1 B1, 2 B3), par « rupture rapide d'un séisme » (1 occurrence correcte), tandis que sa forme plurielle *earthquake ruptures* reçoit les traductions « ruptures de séismes » (2 correctes, 2 B1) et « ruptures des séismes » (1 correcte, 1 B1). Les variantes lexicales *seismic rupture* et *seismic ruptures* sont traduites par « ruptures sismiques » (4 correctes, 3 B1, 1 B3). Ce schéma met en lumière la multiplicité des traductions proposées par le système pour une même famille conceptuelle, ce qui nuit à la cohérence terminologique du texte cible.

VII. Conclusion et perspectives

[Diapositive 44]

Pour conclure, je souhaite souligner plusieurs points :

- L'identification des termes de base et des variantes au sein de chaque texte annoté repose sur une méthode efficace mais chronophage, combinant lecture intégrale, recherche de marqueurs discursifs et recherche par constituants. La vérification systématique des données annotées sur corpus et dans les bases de données terminologiques donne accès aux connaissances du domaine pour confirmer la variation ou vérifier les formes proposées par le système de traduction automatique.
- Le rôle crucial de la vérification sur corpus se confirme : elle permet de vérifier que les termes identifiés correspondent bien à l'usage intégré au discours scientifique et de distinguer les formes attestées des formes erronées produites par le système de traduction automatique.
- Parmi les perspectives de ce travail, nous poursuivons l'annotation des textes restants afin de repérer d'autres types d'erreurs et d'obtenir une vue d'ensemble plus complète

des difficultés rencontrées par le système. Nous envisageons également de comparer les résultats avec ceux d'un système grand public, afin de déterminer si celui-ci propose de meilleures traductions pour les termes et les variantes de notre corpus.

Références

Bawden, R., Bénard, M., Villemonte de La Clergerie, E., Cornejo Cárcamo, J., Dahan, N., Delorme, M., Huguin, M., Kübler, N., Lerner, P., Mestivier, A., Minder, J., Nominé, J.-F., Peng, Z., Romary, L., Tsolakis, P., Zhu, L., & Yvon, F. (2025). MaTOS : Machine Translation for Open Science. *Proceedings of Machine Translation Summit XX: Volume 2*, *Volume 2, Project Presentation Papers*. https://hal.science/hal-05228687

Bénard, M. (2023). Utiliser les syntagmes nominaux complexes anglais pour évaluer la robustesse des systèmes de traduction anglais-français en langue de spécialité. In M. Candito, T. Gerald, & J. G. Moreno (Éds.), *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)* (p. 57-71). ATALA. https://hal.science/hal-04130246

Bénard, M., Kübler, N., Mestivier, A., Minder, J., & Zhu, L. (2024). *Étude des Protocoles d' Évaluation Humaine pour la Traduction de Documents* (p. livrable D4-1.1, 83 pages). Projet ANR MaTOS. https://hal.science/hal-04700009

Biel, Ł. (2023). Variation of legal terms in monolingual and multilingual contexts : Types, distribution, attitudes and causes. In Ł. Biel & H. J. Kockaert (Éds.), Handbook of Terminology : Volume 3. Legal Terminology (p. 90-123). John Benjamins Publishing Company. https://doi.org/10.1075/hot.3.var1

Candel-Mora, M. Á. (2024). Denominative variation in the terminological representation of Women's Health. Cultura, Lenguaje y Representación, 34, 131-150. https://doi.org/10.6035/clr.7888

Carrasco, V. B., & León-Araúz, P. (2023). Denominative variation in the COVID-19 Open Research Dataset corpus. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 29(2), 252-305. https://doi.org/10.1075/term.00071.ben

Collet, T. (2004). What's a term? An attempt to define the term within the theoretical framework of text linguistics. Linguistica Antverpiensia, New Series – Themes in Translation Studies, 3. https://doi.org/10.52034/lanstts.v3i.106

Daille, B. (2017). Term Variation in Specialised Corpora : Characterisation, automatic discovery and applications. John Benjamins. https://doi.org/10.1075/tlrp.19

Delavigne, V. (2020). De l'(in)constance du métalinguistique dans un corpus de vulgarisation médicale. Corela. Cognition, représentation, langage, HS-31. https://doi.org/10.4000/corela.11031

Desmet, I. (2006). Variabilité et variation en terminologie et langues spécialisées : Discours, textes et contextes. Mots, termes et contextes, 235-247.

Fernández-Silva, S. (2016). The cognitive and rhetorical role of term variation and its contribution to knowledge construction in research articles. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 22(1), 52-79. https://doi.org/10.1075/term.22.1.03fer

Fernández-Silva, S. (2019). The Cognitive and Communicative Functions of Term Variation in Research Articles : A Comparative Study in Psychology and Geology. Applied Linguistics, 40(4), 624-645. https://doi.org/10.1093/applin/amy004

Freixa, J. (2002). Variació terminològica : Anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient, La [Ph.D. Thesis, Universitat de Barcelona]. In TDX (Tesis Doctorals en Xarxa). https://www.tdx.cat/handle/10803/1677

Freixa, J. (2022). Chapter 18. Causes of terminological variation. In P. Faber & M.-C. L'Homme (Éds.), Theoretical Perspectives on Terminology : Explaining terms, concepts and specialized knowledge (p. 399-420). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.18fre

Freixa, J., & Fernández-Silva, S. (2017). Terminological variation and the unsaturability of concepts. In P. Drouin, A. Francœur, J. Humbley, & A. Picton (Éds.), Multiple Perspectives on Terminological Variation (p. 155-180). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.18.07fre

Gledhill, C., & Pecman, M. (2018). On alternating pre-modified and post-modified nominals such as aspirin synthesis vs. synthesis of aspirin : Rhetorical and cognitive packing in English science writing. Fachsprache, 40(1-2), 24-46. https://doi.org/10.24989/fs.v40i1-2.1601

Hamon, T., & Nazarenko, A. (2001). Exploitation de l`expertise humaine dans un processus de constitution de terminologie. In D. Maurel (Éd.), Actes de la 8ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs (p. 212-221). ATALA. https://aclanthology.org/2001.jeptalnrecital-long.19/

Kübler, N., Martikainen, H., Mestivier, A., & Pecman, M. (2024). Post-editing neural machine translation in specialised languages : The role of corpora in the translation of phraseological structures. In J. Monti, G. C. Pastor, R. Mitkov, & C. M. Hidalgo-Ternero (Éds.), Recent Advances in Multiword Units in Machine Translation and Translation Technology (p. 57-78). John Benjamins Publishing Company. https://u-paris.hal.science/hal-04909226

Kübler, N., & Pecman, M. (2012). The ARTES bilingual LSP dictionary : From collocation to higher order phraseology (p. 187). Oxford University Press. https://u-paris.hal.science/hal-01134937

León Araúz, P., & Cabezas García, M. (2020). Term and translation variation of multiword terms. MonTI. Monografías de Traducción e Interpretación, 210-247. https://doi.org/10.6035/MonTI.2020.ne6.7

Pecman, M. (2012). Tentativeness in term formation : A study of neology as a rhetorical device in scientific papers. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 18(1), 27-58. https://doi.org/10.1075/term.18.1.03pec

Pecman, M. (2014). Variation as a cognitive device : How scientists construct knowledge through term formation. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 20(1), 1-24. https://doi.org/10.1075/term.20.1.01pec

Pecman, M. (2021). 10th anniversary of the ARTES terminological and phraseological database. Izdavalacko izdanje Instituta za hrvatski jezik i jezikoslovlje (Publication of the Institut of Croatian language and linguistics). https://u-paris.hal.science/hal-03175473

Peng, Z., Bawden, R., & Yvon, F. (2025). *Investigating Length Issues in Document-level Machine Translation* (arXiv:2412.17592). arXiv. https://doi.org/10.48550/arXiv.2412.17592

Tkachenko, M., Malyuk, M., Holmanyuk, A., & Liubimov, N. (2020). Label Studio : Data labeling software [Logiciel]. https://github.com/HumanSignal/label-studio

Seghezzi, N. (2011). Variación terminológica y canal de comunicación. Estudio contrastivo de textos especializados escritos y orales sobre lingüística [Ph.D. Thesis, Universitat Pompeu Fabra]. In TDX (Tesis Doctorals en Xarxa). https://www.tdx.cat/handle/10803/52066

Suárez de la Torre, M. (2004). Análisis contrastivo de la variación denominativa en textos especializados : Del texto original al texto meta [Ph.D. Thesis, Universitat Pompeu Fabra]. In TDX (Tesis Doctorals en Xarxa). https://www.tdx.cat/handle/10803/7495