Rapport du groupe de travail "Traductions et science ouverte"

Susanna Fiorini, Franck Barbin, Martine Garnier-Rizet, Katell Hernandez Morin, Franziska Humphreys, Amélie Josselin-Leray, Natalie Kübler, Rudy Loock, Hanna Martikainen, Jean-François Nominé, et al.

Rapport du groupe de travail Traductions et science ouverte

Novembre 2020

Groupe de travail

Pilote

Susanna FIORINI (Traductrice et consultante en communication multilingue)

Membres

Franck BARBIN (Université Rennes 2 / LIDILE)
Martine GARNIER-RIZET (ANR)
Katell HERNANDEZ MORIN (Université Rennes 2 / LIDILE)
Franziska HUMPHREYS (EHESS)
Amélie JOSSELIN-LERAY (Université de Toulouse Jean Jaurès / CLLE)
Natalie KÜBLER (Université de Paris / CLILLAC-ARP)
Rudy LOOCK (Université de Lille / STL)
Hanna MARTIKAINEN (École Supérieure d'Interprètes et de Traducteurs / CLESTHIA)
Jean-François NOMINÉ (Institut de l'information scientifique et technique/CNRS)
Cornelia PLAG (Université de Coimbra / OPERAS)
Caroline ROSSI (Université Grenoble Alpes / ILCEA4)
François YVON (LIMSI / CNRS)

Remerciements

Le groupe de travail remercie les expert·es, les plateformes de services et de contenus, les éditeurs et les chercheurs·euses ayant contribué aux réflexions présentées dans ce rapport.
Par ordre alphabétique :
Cairn.info
Caroline Champsaur (OCDE)
Éditions Quæ
Érudit
HAL
Huma-Num
John Libbey Eurotext
Lauréats de l'appel à projets Traductions scientifiques
Lynne Bowker (Université d’Ottawa)
Nicolas Bacaër (Institut de recherche pour le développement)
OpenEdition
Sharon O’Brien (Dublin City University)
Sheila Castilho (ADAPT Centre - Dublin City University)

Index

Résumé

Selon l'*Initiative d'Helsinki sur le multilinguisme dans la communication savante*, celui-ci permet de continuer à mener des recherches pertinentes au niveau local, de créer de l'impact par la diffusion des résultats de la recherche dans sa propre langue, de valoriser la diversité des travaux scientifiques et d'interagir avec la société. Or, si la culture scientifique est majoritairement véhiculée par une seule langue, partager des connaissances au-delà des organismes de recherche et des universités devient difficile. Le contraire de l'esprit de la science ouverte dont l'un des principes fondamentaux est la démocratisation de l'accès au savoir produit par la recherche. L'Initiative d'Helsinki a servi de point d'ancrage de la démarche du groupe de travail « Traductions et science ouverte » pour son rapport.

Pour les membres du groupe, la traduction constitue clairement une option possible pour répondre à cette nécessité d'ouverture. Leur objectif est d'identifier des possibilités techniques pour développer la diffusion multilingue de la science, en exploitant les récents progrès des technologies de la traduction. Ainsi, le multilinguisme à une large échelle dans la communication scientifique sera favorisé ; les chercheurs pourront publier dans la langue de leur choix sans pour autant être pénalisés ; un nouveau modèle d'accès, universel et multilingue, à l'information scientifique, verra le jour. Mais une condition est indispensable : « l'humain doit rester au cœur du processus, les technologies devant optimiser le travail sans devenir une contrainte ou une source de frustration pour les utilisateurs, que ce soit les intervenants dans le processus de traduction ou les lecteurs finaux. »

Les membres du groupe attirent également l'attention sur trois éléments essentiels à la réussite du multilinguisme :

- une adaptation de l'écosystème de l'édition scientifique ;
- des actions politiques pour repenser les systèmes et les métriques d'évaluation, ainsi que les mécanismes de financement ;
- un changement d'ordre culturel chez les universitaires, les chercheurs et les enseignants-chercheurs afin que la valeur des publications non anglophones soit pleinement reconnue.

Le rapport propose de poursuivre une double ambition : 1) favoriser un rayonnement de la production scientifique en français vers d'autres langues dans tous les continents et 2) briser les barrières linguistiques pour les citoyens, organisations et entreprises francophones souhaitant accéder aux résultats de la recherche internationale.

Dans un premier temps, le groupe montre la nécessité d'une approche raisonnée de la traduction, au vu de la masse de publications, pour tenir compte entre autres des usages et des besoins disciplinaires. Une telle démarche différenciée apparaît nécessaire dans l'optique de valorisation de la production scientifique française à l'international. Dans ce but, des expérimentations technologiques seront réalisées dans cinq disciplines scientifiques, sur trois paires linguistiques (français → anglais, anglais → français et français → espagnol) pour des raisons de disponibilité de ressources linguistiques et d'audience mondiale importantes, et sur des formats de publication abordables, notamment les métadonnées, les résumés et les comptes rendus d'ouvrages.

Le groupe présente ensuite un inventaire des outils de traduction, constitué de deux sections : une dédiée aux outils de traduction automatique et une consacrée aux outils de traduction assistée par ordinateur (TAO).

Enfin, le groupe de travail recommande à court et moyen terme les actions et les expérimentations listées ci-après :

1. Analyse de la nature et de la volumétrie des corpus multilingues identifiés et étude de nouvelles possibilités pour la collecte ;
2. Traitement des corpus collectés afin d'obtenir des bases de test et d'apprentissage et des ressources linguistiques mutualisées pour alimenter les systèmes de traduction ;
3. Évaluation de moteurs de traduction automatique en utilisant les bases de test et d'apprentissage ;
4. Organisation de journées d'étude rassemblant les porteurs des projets lauréats de l'appel Traductions scientifiques et d'autres acteurs pertinents pour faire des propositions sur la base de constats et résultats concrets ;
5. Création d'un démonstrateur pour préfigurer un processus de traduction à grande échelle dont les résultats pourraient être révisés en vue d'être adaptés et pris en charge dans une chaîne éditoriale ;
6. Élaboration d'un guide à destination des chercheurs et des institutions de recherche sur la traduction automatique, la rédaction en langue étrangère et la « rédaction claire » (adaptée à la traduction automatique) ;
7. Étude de pistes de collaboration dans les réseaux d'éditeurs européens pour la constitution de corpus.

Les travaux du groupe s'inscrivent dans le cadre du Comité pour la science ouverte (CoSO) en partenariat avec la Délégation générale à la langue française et aux langues de France (DGLFLF), et se placent dans la continuité de l'appel à projets *Traductions scientifiques*, lancé en 2018 par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI).
Le groupe est suivi par un comité de pilotage constitué de représentants des institutions partenaires (MESRI, CoSO, DGLFLF).

1. Multilinguisme et science ouverte

Depuis l'intensification de l'activité scientifique dans l'après-guerre et, plus tard, avec l'avènement de l'Internet, l'anglais s'est affirmé mondialement en tant que langue de la communication technique et scientifique. Dès les années 1960, par exemple, la langue française a souffert d'un déficit lexical de milliers de mots supplémentaires chaque année au regard de la progression enregistrée pour l'anglais[1].

Ce rôle de *lingua franca* s'est progressivement consolidé dans le milieu académique, à la fois dans la recherche et dans l'enseignement supérieur. À titre d'exemple, 83,7% des 23 millions de documents scientifiques rassemblés dans le cadre de la création de l'archive électronique ISTEX étaient en 2019 en anglais[2].

Si elle a le mérite de favoriser les échanges dans un contexte scientifique de plus en plus internationalisé, cette hégémonie linguistique est génératrice d'inégalité dans l'accès au système de publication et limite la diffusion des connaissances scientifiques au sein des sociétés des pays non anglophones.

Afin de publier dans les revues à impact élevé et ainsi augmenter la visibilité de leur travail, les chercheurs dans certaines disciplines sont en effet tenus de publier en anglais, ce qui a un impact considérable sur leur carrière ; outre les difficultés de rédaction en langue étrangère, qui limitent la richesse de la pensée et les capacités d'expression de certains concepts, des chercheurs non anglophones considèrent également qu'ils sont pénalisés car certains relecteurs se concentrent davantage sur leur niveau d'anglais que sur la qualité des résultats scientifiques et la logique de l'exposé[3]. Par ailleurs, on constate que les relecteurs peuvent être gênés dans leur évaluation par des modes d'argumentation non anglophones. Ainsi, Lillis et Curry relèvent la remarque d'un relecteur : « Il y a des formulations qui, d'après moi, sont un peu exagérées et trop prétentieuses. (…) Ce n'est peut-être pas la langue, mais c'est juste un peu trop latin pour un Européen du Nord. »[4] Selon une récente étude, la publication en langue étrangère imposée aux chercheurs peut enfin générer des problèmes de rédaction et compréhension en lecture, de l'anxiété, voire des coûts supplémentaires[5].

L'actuel système à dominante anglophone entraîne également des difficultés pour les chercheurs qui, du fait des usages propres à leurs disciplines, notamment dans les sciences humaines et sociales, ont la possibilité de rédiger leurs publications dans leur langue maternelle. Dans ce cas, le problème est lié à un manque de visibilité des travaux rédigés dans des langues autres que l'anglais, souvent moins bien référencés dans les principaux

---

[1] L. Bowker, J. Ciro, 2019, Machine translation and global research: Towards improved machine translation literacy in the scholarly community, Bingley, UK: Emerald Publishing

[2] ISTEX - Socle de la bibliothèque scientifique numérique nationale [consulté le 29 octobre 2020]

[3] L. Bowker, J. Ciro, 2019, Machine translation and global research: Towards improved machine translation literacy in the scholarly community, Bingley, UK: Emerald Publishing

[4] Lillis, Theresa and Curry, Mary Jane, 2010, Academic Writing in a Global Context: The politics and practices of publishing in English. Abingdon: Routledge

[5] V. Ramírez-Castañeda, 2020, Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. PLoS ONE 15(9): e0238372. https://doi.org/10.1371/journal.pone.0238372

référentiels disciplinaires internationaux. Une autre étude[6] démontre que le fait de publier dans une langue autre que l'anglais est considéré comme l'indice d'une mauvaise qualité de la recherche et de localisme, une idée largement partagée dans le monde de la recherche.

D'autre part, une culture scientifique majoritairement véhiculée par la langue anglaise ne favorise pas la démocratisation de l'accès au savoir produit par la recherche, l'un des principes fondamentaux de la science ouverte. Dans le cadre de la crise sanitaire liée à la pandémie de Covid-19, qui a fait apparaître le besoin d'une diffusion des connaissances scientifiques et médicales auprès des citoyens, un groupe d'enseignants-chercheurs a écrit que « si la science fait l'objet exclusivement d'une communication en anglais, elle risque de ne pas répondre pleinement à sa troisième mission, celle d'informer les citoyens dans leurs langues maternelles. »[7]

Dans ce contexte, plusieurs initiatives ont vu le jour afin de rappeler l'importance du multilinguisme dans la communication scientifique. Parmi celles-ci, l'Initiative d'Helsinki sur le multilinguisme, point d'ancrage important de la démarche de ce groupe de travail, propose des recommandations[8] à destination des tous les acteurs pouvant promouvoir le changement : décideurs, dirigeants, universités, instituts de recherche, bailleurs de fonds pour la recherche, bibliothèques et chercheurs. Les initiateurs de l'Initiative expliquent qu'il ne s'agit pas « d'une question de nationalisme ou d'aversion à l'égard de la langue anglaise » mais plutôt de « développer un système de publication qui réponde aux besoins des chercheurs et de leur public, et qui valorise la diversité des travaux scientifiques. »[9]

Afin de permettre aux chercheurs de publier dans la langue de leur choix sans pour autant être pénalisés, et de créer un nouveau modèle d'accès, universel et multilingue, à l'information scientifique, la traduction constitue clairement une option possible. Toutefois, aux moins trois défis se posent :

- les universités, les instituts de recherche et les laboratoires manquent souvent de ressources à consacrer à la traduction ;
- il n'est pas toujours facile de trouver des traducteurs experts capables de traduire des textes hautement spécialisés dans des délais raisonnables ;
- d'un point de vue global, les contenus de communication scientifique sont d'un volume tel qu'il serait impossible de couvrir les besoins de traduction par les moyens humains professionnels disponibles, experts ou non.

Afin de pallier ces difficultés et de favoriser le multilinguisme à une large échelle dans la communication scientifique, une nouvelle voie d'exploration semble s'ouvrir. Les récents progrès techniques laissent envisager la possibilité de recourir à des technologies de la traduction de plus en plus performantes. L'environnement de travail des professionnels de la traduction ne cesse en effet de s'enrichir d'outils informatiques ayant vocation à optimiser le processus de traduction : logiciels de traduction assistée par ordinateur, bases

---

[6] Di Bitetti, Mario S., and Julián A. Ferreras, 2017, Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications, Ambio 46.1: 121-127

[7] Z. Taşkın, G. Doğan, E. Kulczycki, A. Zuccala, 2020, Science needs to inform the public. That can't be done solely in English, LSE blog [consulté le 17 août 2020]

[8] Initiative d'Helsinki sur le multilinguisme dans la communication savante [consulté le 10/11/2020]

[9] Emanuel Kulczycki, Henriikka Mustajoki, Janne Pölönen, Vidar Røeggen, 2019, Polyglots need protection, Research Europe [consulté le 10/11/2020]

terminologiques[10], mémoires de traduction[11], systèmes de gestion de projets collaboratifs, moteurs de traduction automatique et interfaces de post-édition[12], outils de contrôle qualité[13], etc. Une condition indispensable doit cependant être respectée pour que ces technologies puissent réellement apporter une aide à la traduction : l'humain doit rester au cœur du processus, les technologies devant optimiser le travail sans devenir une contrainte ou une source de frustration pour les utilisateurs, que ce soit les intervenants dans le processus de traduction ou les lecteurs finaux. L'impact et la pertinence de ces outils doivent donc être évalués au cas par cas afin de mettre en place des solutions plus ou moins informatisées et automatisées selon les contextes d'usage.

L'objectif du groupe de travail est donc d'identifier des possibilités dans ce sens afin d'initier une montée en charge de la traduction de la production scientifique en s'appuyant sur les technologies de la traduction, en particulier sur les outils de traduction automatique.
Ce rapport présente un premier aperçu de l'état de l'art des technologies de la traduction, des bonnes pratiques d'usage et des pistes d'action, à court et moyen terme, afin d'optimiser les processus de traduction dans le but notamment de :

- favoriser la visibilité internationale des publications scientifiques en langue française, en particulier pour les sciences humaines et sociales
- briser la barrière que représente l'usage dominant de l'anglais pour la diffusion large des connaissance scientifiques dans les sociétés des pays non anglophones
- proposer des solutions pour lutter contre les inégalités subies par les chercheurs non anglophones

Les travaux du groupe s'inscrivent dans le cadre du Comité pour la science ouverte (CoSO) en partenariat avec la Délégation générale à la langue française et aux langues de France (DGLFLF), et se placent dans la continuité de l'appel à projets Traductions scientifiques, lancé en 2018 par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI). Le groupe est suivi par un comité de pilotage constitué de représentants des institutions partenaires (MESRI, CoSO, DGLFLF).

Avant d'analyser les possibilités que ces technologies pourront offrir, il convient néanmoins de noter que le multilinguisme ne pourra s'installer durablement dans l'édition scientifique que dans un écosystème adapté. Il ne suffira pas de déployer des technologies pour optimiser les processus de traduction sans une vraie stratégie de médiation locale et de valorisation internationale des contenus multilingues ; cela devra passer par les référentiels disciplinaires internationaux, par des réseaux de médiation et de valorisation spécifiques aux différents publics (local ou international, académique ou généraliste), par la collaboration avec des acteurs clés, comme les bibliothécaires, les documentalistes ou encore les journalistes scientifiques. Très souvent, en effet, les difficultés d'accès à l'information scientifique ne sont

[10] Bases de données contenant des entrées terminologiques et des informations associées. La plupart des bases terminologiques sont multilingues et contiennent des données terminologiques dans plusieurs langues.
[11] Bases de données qui enregistrent les phrases, paragraphes ou segments de texte traduits pour réutilisation ultérieure.
[12] Activité consistant à réviser et corriger le texte brut pré-traduit automatiquement par un moteur afin d'atteindre le niveau de qualité souhaité.
[13] Fonctionnalités des outils de TAO permettant de vérifier automatiquement des éléments de la traduction : terminologie, chiffres, balises, ponctuation, incohérences, etc.

pas exclusivement liées aux barrières linguistiques, mais aussi aux problématiques de référencement, découvrabilité et compréhensibilité conceptuelle des contenus pour les différents types de publics.

Des actions politiques seront par ailleurs nécessaires pour repenser les systèmes et les métriques d'évaluation, ainsi que les mécanismes de financement, qui aujourd'hui favorisent la publication en anglais au détriment de la production dans d'autres langues. Les chercheurs devront être encouragés à publier et à traduire des contenus dans des langues autres que l'anglais grâce à une meilleure reconnaissance de ces travaux et à des évaluations basées sur la qualité effective des publications et non pas uniquement sur leur facteur d'impact, pour un accès plus égalitaire aux financements. Ceci implique également un changement d'ordre culturel chez les universitaires afin que la valeur des publications non anglophones soit pleinement reconnue par les comités de sélection, mais aussi lors des évaluations de l'HCERES et du CNU, par exemple.

Une plus grande ouverture des publications et la généralisation de licences ouvertes permettraient enfin de favoriser la traduction et la circulation des contenus multilingues, ainsi que de récupérer plus facilement des ressources linguistiques, nécessaires au développement de solutions ouvertes et maîtrisées par les communautés scientifiques - l'objectif étant également de ne pas subir la concentration des acteurs commerciaux du secteur des technologies de la traduction, de préserver les capacités d'initiative sur les contenus scientifiques, et d'alimenter des environnements plus ouverts (bases d'apprentissage, algorithmes et logiciels). Le besoin d'ouverture ne concerne donc pas seulement les contenus et les ressources linguistiques, mais aussi les outils technologiques, en particulier les moteurs de traduction automatique.

Afin d'atteindre tous ces objectifs qui touchent également la sphère politique, il apparaît nécessaire de créer des collaborations à l'échelle internationale. À ce propos, des initiatives se multiplient et se développent en Europe et ailleurs ; à noter, outre l'Initiative d'Helsinki sur le multilinguisme déjà mentionnée, le Groupe de travail sur le Multilinguisme du réseau OPERAS[14], la San Francisco Declaration on Research Assessment[15], le projet Triple[16], ou encore des initiatives à échelle éditoriale, comme celles des revues Target[17] et Handbook of Translation Studies[18]. Le groupe de travail suivra ces initiatives afin d'établir toute connexion pertinente. Par exemple, le Groupe de travail sur le Multilinguisme OPERAS, créé pour promouvoir les pratiques de traduction, les outils de découverte multilingues et l'utilisation des langues nationales dans l'édition en sciences humaines et sociales, envisage de développer une plateforme collaborative afin de mettre en contact éditeurs, chercheurs et traducteurs et favoriser ainsi la traduction de publications. Même si cette démarche est surtout orientée à la mise en relation des différents acteurs, et moins sur le déploiement des technologies de la traduction, il sera intéressant d'étudier d'éventuels points de contact et collaboration.

[14] Multilingualism WG – OPERAS
[15] DORA – San Francisco Declaration on Research Assessment
[16] Transforming Research through Innovative Practices for Linked interdisciplinary Exploration
[17] Target. International Journal of Translation Studies
[18] Handbook of Translation Studies Online

2. Travail du groupe

2.1 Périmètre disciplinaire

Des études[19] montrent que les résultats de la recherche n'intéressent pas que les membres de communautés disciplinaires restreintes, mais également des « lecteurs inattendus ». Cette forme de lecture, favorisée par la publication de contenus en accès ouvert et généralement liée à des sujets d'actualité, concerne de nombreuses disciplines. La crise sanitaire liée à la pandémie de Covid-19 constitue un exemple particulièrement parlant : dès les premiers mois de 2020, de nombreux lecteurs ont voulu consulter des articles médico-scientifiques afin de repérer des informations fiables dans la vague de désinformation circulant dans la presse généraliste et les réseaux sociaux. Or, près de 85% des articles sur la maladie à coronavirus référencés dans la base de données de l'OMS ont été publiés en anglais[20], ce qui n'a pas non plus facilité l'accès à l'information pour les professionnels de santé du monde entier.

Ces arguments ne doivent pas être lus comme un appel à tout traduire de manière indiscriminée, mais plutôt à développer une approche raisonnée de la traduction tenant compte des usages et des besoins disciplinaires, du niveau d'accessibilité des contenus pour les différents publics ainsi que de la typologie des textes. Une telle démarche différenciée apparaît également nécessaire dans l'optique de valorisation de la production scientifique française à l'international.

S'agissant d'une analyse complexe et de longue haleine, le groupe de travail et le comité de pilotage ont proposé une première sélection de disciplines selon les critères suivants : la volumétrie de publication en France et à l'étranger, le taux de publication en accès ouvert[21], la compatibilité du langage et du style d'écriture disciplinaire avec la traduction humaine, semi-automatique et automatique, l'intérêt et le niveau d'accessibilité des contenus pour un lectorat savant et généraliste, les ressources linguistiques disponibles (corpus et bases terminologiques), l'existence de besoins constatés au sein des communautés disciplinaires.

Les expérimentations démarreront donc dans les disciplines ci-dessous. Des domaines et sous-domaines seront identifiés en fonction de la pertinence des projets, des acteurs et des ressources disponibles.

2.1.1 Archéologie

L'archéologie suscite depuis toujours l'intérêt du grand public, curieux de ses résultats et de ses méthodes. Des revues phares françaises, d'audience nationale et internationale, ont récemment engagé des démarches d'internationalisation de leurs contenus afin d'accroître la visibilité de la recherche française en archéologie dans le monde. Parmi celles-ci, le *Bulletin de la Société Préhistorique Française*, la revue *ArchéoSciences*, et bien d'autres. Cette

[19] Par exemple, Umberto, le détecteur de lecteurs d'OpenEdition [consulté le 09/11/2020]
[20] Z. Taşkın, G. Doğan, E. Kulczycki, A. Zuccala, 2020, Science needs to inform the public. That can't be done solely in English, LSE blog [consulté le 17 août 2020]
[21] Les données relatives aux publications ayant au moins un auteur avec une affiliation française sont disponibles en ligne sur le site du Baromètre de la science ouverte du Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation

importante production scientifique devrait donc permettre de récupérer des ressources linguistiques existantes afin de les exploiter dans le cadre de projets d'expérimentation.

2.1.2 Géographie

Les contenus de cette discipline peuvent intéresser le lectorat académique (les chercheurs mais également les étudiants de nombreuses spécialités) et le grand public. Plusieurs revues françaises et européennes (*Cybergeo*, *Via.Tourism Review*, *Brussels Studies*, etc.) fonctionnent déjà avec un modèle multilingue, ce qui devrait permettre d'accéder à des ressources linguistiques. Par ailleurs, le style d'écriture factuel de la discipline se prête plutôt bien à la traduction, même si le champ lexical, spécifique et en même temps très varié, pose un défi concernant la gestion de la terminologie.

2.1.3 Médecine

La médecine est une discipline fondamentale pour la société. Or, la plupart des résultats sont publiés en anglais, rendant *de facto* l'information médicale, plus ou moins spécialisée, inaccessible à une partie des professionnels et citoyens. Le besoin de multilinguisme est donc évident, au point que plusieurs expérimentations ont déjà été menées en France et à l'international. L'utilisation des technologies de la traduction semble être particulièrement pertinente : le public cible apprécie en effet le style rédactionnel résultant de l'intervention de la machine, et la post-édition par des spécialistes du domaine est prometteuse. Par ailleurs, compte tenu de la volumétrie de la production scientifique médicale française, des solutions technologiques permettraient de valoriser systématiquement les résultats rédigés en français à échelle internationale.

2.1.4 Économie

Intéressant plusieurs domaines d'études et de recherche, et donc un grand nombre de chercheurs et étudiants, la production scientifique en économie est dominée par l'anglais. Il apparaît donc pertinent d'envisager rapidement des solutions en faveur du multilinguisme, d'autant plus que le langage disciplinaire, plus factuel et moins conceptuel par rapport à d'autres disciplines, peut bien se prêter à la traduction automatique. Cette discipline présentant par ailleurs un fort lien avec l'actualité, ses résultats sont susceptibles d'intéresser le plus grand nombre.

2.1.5 Sciences de la terre, de l'environnement et de la planète (Géosciences)

Cette dénomination regroupe plusieurs sous-domaines susceptibles de concerner à la fois un lectorat académique, d'enseignants du secondaire, d'ingénieurs et un public généraliste. En effet, la volcanologie, la sismologie et la glaciologie par exemple, outre leur intérêt scientifique, constituent des domaines dans lesquels l'information scientifique est indispensable pour les décideurs publics (menaces d'éruptions, de tremblements de terre, d'avalanches, etc.), mais aussi pour les populations. De même, les questions de la vie sur d'autres planètes et de l'histoire sismique de celles-ci (cf. la mission InSight sur Mars en 2018) rencontrent un grand intérêt chez le public non expert. Par ailleurs, les sciences de la terre, de l'environnement et de la planète publient presque exclusivement en anglais. Les besoins d'aide auprès des chercheurs sont donc importants. En outre, depuis que le Bulletin de la Société Géologique de Paris ne publie plus en français, le besoin de diffuser la recherche française en français est devenu de plus en plus criant.

Il apparaît donc pertinent d'étudier des solutions favorisant le multilinguisme dans ce domaine. Des ressources linguistiques devraient être accessibles grâce aux publications en accès ouvert (environ 50% de la production française, selon le Baromètre de la Science Ouverte) et aux référentiels constitués par les universités. L'institut de Physique du Globe de Paris, qui fait partie de l'Université de Paris, peut aider à fournir des ressources.

2.1.6 D'autres disciplines à étudier à moyen terme

Compte tenu de la richesse de la production scientifique française et des enjeux liés à l'ouverture de la science à la société, il sera important de poursuivre ces réflexions afin d'élargir les dispositifs expérimentaux au plus grand nombre de disciplines jugées pertinentes.

Par exemple, les disciplines choisies pour cette première phase d'expérimentation ne représentent pas toute la diversité et la richesse du langage des sciences humaines et sociales, dont les résultats sont traditionnellement publiés en français et donc plus difficilement valorisés à l'international. Le groupe préconise donc d'inclure à moyen terme d'autres disciplines, comme la philosophie, la psychologie ou la linguistique.

Des réflexions en matière de multilinguisme et des expérimentations de traduction automatique à petite échelle sont par ailleurs déjà en cours dans les domaines des sciences expérimentales et mathématiques, ou dans le Traitement Automatique des Langues. Il sera intéressant de les étudier également sur le moyen terme.

2.2 Périmètre linguistique

Les travaux du groupe ont une double ambition : favoriser un rayonnement de la production scientifique en langue française dans tous les continents et briser les barrières linguistiques pour les citoyens, organisations et entreprises francophones souhaitant accéder aux résultats de la recherche internationale. Ce travail pourra par ailleurs bénéficier à d'autres communautés linguistiques et favoriser des collaborations hors de la sphère francophone.

Pour atteindre ces objectifs à grande échelle, il est nécessaire d'élargir et d'optimiser les processus de traduction dans le plus grand nombre de langues. Ce travail d'envergure requiert néanmoins une stratégie de priorisation et planification.

Afin d'établir une première sélection de langues à étudier pour chaque discipline, les critères suivants ont été considérés : disponibilité de ressources linguistiques, maturité des technologies existantes, audience potentielle et pertinence disciplinaire.

Les premières expérimentations porteront donc sur les paires linguistiques suivantes :

- français → anglais
- anglais → français
- français → espagnol

Pour la suite des travaux à moyen terme, d'autres paires linguistiques apparaissent particulièrement pertinentes :

- français ↔ arabe : des expérimentations actuellement en cours devraient permettre d'exploiter des ressources linguistiques afin d'améliorer les technologies existantes dans certaines disciplines.
- français → chinois, français → portugais : la traduction des publications rédigées en français vers ces deux langues devrait permettre d'augmenter considérablement leur impact. Un autre argument concernant les langues asiatiques est que celles-ci sont en général peu parlées par les chercheurs, donc la qualité des traductions de travaux francophones réalisées dans ces pays est difficilement vérifiable. En mettant à disposition des outils et des processus optimisés, la traduction de travaux pertinents vers les langues asiatiques pourrait être systématisée à la source.
- français ↔ allemand, français ↔ italien : de nombreux liens et collaborations académiques existent dans plusieurs disciplines entre ces trois pays voisins. Cela devrait permettre d'obtenir facilement des ressources linguistiques pour développer des dispositifs de traduction favorisant davantage les échanges et la circulation des savoirs.

2.3 Périmètre documentaire

La publication des résultats de la recherche scientifique se fait désormais sous plusieurs formes selon des usages et des besoins propres à chaque discipline : livres, articles de revues, actes de colloques, carnets de recherche et annonces scientifiques, auxquels sont toujours associées des métadonnées (titres, résumés et mots-clés) permettant d'identifier, diffuser et valoriser chaque publication. Afin de déployer des dispositifs de traduction humaine, automatique ou semi-automatique, il convient donc d'analyser les contraintes, usages et besoins liés à chaque type de contenu. Par exemple, la qualité d'une traduction automatique varie considérablement selon la longueur du texte, la complexité du langage, le style d'écriture et les mécanismes d'argumentation ; on peut donc espérer obtenir de meilleurs résultats en utilisant des outils de traduction automatique sur un résumé ou sur un compte rendu, plutôt que sur un article ou un ouvrage dans son intégralité.

Les expérimentations porteront donc en premier lieu sur des formats abordables, notamment les métadonnées, les résumés et les comptes rendus d'ouvrages, mais seront également inclus des formats plus longs et complexes, comme des articles *in extenso* ainsi que des chapitres d'ouvrages. Des réflexions ultérieures seront prévues à moyen terme afin d'étudier les possibilités pour les autres types de publications.

2.4 Besoins et pratiques de traduction

Identifier tous les besoins et les pratiques de traduction existant dans le milieu de la communication scientifique est une tâche très complexe. Il s'agit en effet d'un paysage très varié, déterminé par plusieurs facteurs : usages et contraintes disciplinaires, moyens et ressources disponibles, enjeux éditoriaux, etc. Il apparaît donc compliqué de pouvoir proposer un scénario universel, adapté à tous les contextes.

Il est cependant possible de s'interroger sur une première question cruciale : à quel moment du cycle de vie d'une publication est-il le plus pertinent de situer le processus de traduction ? Pour répondre à cette question, trois moments principaux ont été identifiés :

1. Pré-production : phase de rédaction
2. Production : phase d'édition
3. Post-production : après la publication

Parmi les chercheurs-auteurs, il semblerait que l'activité de traduction porte essentiellement sur les résumés et dans une moindre mesure sur les textes intégraux. Prévoir un processus de traduction au niveau de la pré-production pourrait signifier de demander à l'auteur de rédiger ou éditer une publication dans sa langue maternelle et de la traduire dans d'autres langues suffisamment maîtrisées à l'aide d'outils de traduction et de rédaction. Une telle stratégie permettrait de bénéficier de connaissances disciplinaires, et donc terminologiques et phraséologiques, indispensables dans l'activité de post-édition de traduction automatique. Mais *quid* des autres compétences nécessaires pour mener à bien un travail de post-édition ? La post-édition est en effet une activité nécessitant une excellente maîtrise des langues de travail et des technologies de la traduction (TAO et traduction automatique), ainsi qu'un entraînement spécifique pour acquérir les bons réflexes. Certes, il est possible de mettre en place des formations spécifiques sur la traduction automatique et la post-édition à destination des chercheurs-auteurs. Selon une enquête récemment menée par le groupe de travail sur le multilinguisme du réseau OPERAS, 66% des chercheurs-auteurs sondés ont déclaré être intéressés à suivre des formations pour mieux connaître ces outils et processus[22]. Cependant, il faudra considérer que, la traduction ne faisant pas naturellement partie de leurs missions et ayant déjà de nombreuses tâches annexes chronophages, les chercheurs-auteurs manqueront sans doute de temps à consacrer à cet apprentissage et à l'activité de traduction ou post-édition sur la durée. Par exemple, on demande déjà aux auteurs de produire des résumés, et le plus souvent des traductions de résumés, de leurs publications ; or, ce travail est parfois fait rapidement à l'issue de la rédaction de l'article ou de l'ouvrage intégral. Compte tenu de cela, il n'est pas certain que les chercheurs-auteurs acceptent de bon gré de réaliser une post-édition de traduction automatique du texte intégral tout juste rédigé, d'autant plus si cela est demandé de manière systématique. Avant de construire des scénarios de travail dans ce sens, il apparaît donc nécessaire d'interroger, à assez grande échelle et sur base disciplinaire, le public intéressé et de le sensibiliser davantage sur l'importance de la traduction dans l'édition scientifique[23]. Dans tous les cas, il faudra être en mesure de mettre à disposition des technologies adaptées en termes de performance, accessibilité et ergonomie afin que le travail de post-édition soit le plus fluide et rapide possible.

Une autre limite de ce mode de fonctionnement serait la contrainte imposée par les connaissances linguistiques des chercheurs qui, dans les meilleurs des cas, maîtrisent une ou deux langues supplémentaires, en plus de leur langue maternelle, à un niveau suffisant pour effectuer une post-édition de traduction automatique. En outre, maîtriser une langue seconde ne signifie pas seulement maîtriser la syntaxe, la terminologie et la phraséologie. C'est aussi une question de langue-culture et de mode de pensée, et par conséquent, de mode d'argumentation. Or, maîtriser le mode d'argumentation de l'article scientifique dans

[22] D. Leão, 2020, The Survey "Multilingualism in Scholarly Communication": a Preliminary Report, Atelier sur le multilinguisme de la conférence OASPA 2020

[23] Voir par exemple : Patrik Lambert, Jean Senellart, Laurent Romary, Holger Schwenk, Florian Zipser, et al., Collaborative Machine Translation Service for Scientific texts. Proc. of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2012, Avignon, France. pp.11-15. hal-00723349f

des langues autres que sa langue maternelle demande d'excellentes connaissances linguistiques et une conscience linguistique très affutée pour améliorer un texte en post-édition sur ce plan. Dans le but de développer le multilinguisme le plus largement possible et donc d'élargir les dispositifs au plus grand nombre de langues et publications, il sera de toute manière nécessaire de réfléchir à d'autres solutions complémentaires.

Un processus de traduction en phase de production serait en revanche confié à la maison d'édition qui, en centralisant le travail, pourrait assurer un rôle de garant de la cohérence et de la qualité linguistique des publications traduites. Une stratégie de ce type, envisagée par certaines équipes lauréates de l'appel *Traductions scientifiques*, demande néanmoins des ressources et des moyens financiers durables, des équipes de traducteurs spécialisés et fidélisés ainsi que des outils technologiques adaptés. En réunissant ces conditions, il serait possible de mettre en place un processus de traduction augmentée avec un ou plusieurs moteurs de traduction automatique intégrés dans un outil de TAO. L'utilisation de la TAO permettrait de partager des ressources linguistiques, d'optimiser la gestion terminologique et l'alimentation de mémoires de traduction qui serviraient de bases d'apprentissage pour la traduction automatique, en instaurant donc un cercle vertueux. Mais, encore une fois, la mise en place d'un tel processus nécessite des connaissances métier et des efforts de gestion ; il serait donc souhaitable que des postes de coordination des traductions soient créés au sein des équipes éditoriales.

Une telle approche pourrait également être adoptée en phase de post-production, avec une centralisation des traductions assurée par des acteurs opérant à plus grande échelle, comme par exemple les plateformes de diffusion de contenus ou d'autres grandes infrastructures. C'est le cas aujourd'hui de la plateforme de revues et ouvrages en sciences humaines et sociales Cairn.info pour sa version internationale[24]. Cette stratégie permettrait de faciliter le déploiement d'un seul modèle pour plusieurs éditeurs disciplinaires, pour une mutualisation des coûts, des ressources et des bonnes pratiques. Clairement, cela demanderait une organisation et des moyens encore plus importants afin de pouvoir obtenir, à plus grande échelle, des textes traduits de qualité publiable. Ces moyens importants seraient néanmoins plus soutenables à l'échelle d'une plateforme de diffusion qu'à l'échelle de chaque éditeur pris isolément.

Il faut cependant noter que l'objectif d'une qualité publiable, en termes linguistiques et éditoriaux, n'est pas la seule voie possible. Certaines plateformes internationales, telles que EBSCO ou Taylor & Francis Online, se sont en effet dotées de moteurs de traduction automatique intégrés afin de proposer aux lecteurs un service de traduction instantanée sans supervision humaine. Signalées par des mentions explicites indiquant la nature automatique des traductions, celles-ci permettent d'avoir une idée des contenus du texte sans avoir la présomption de vouloir remplacer des traductions ou des post-éditions humaines. Ce type d'usage (appelé *gisting* en anglais, ou traduction informative en français) est de plus en plus répandu dans des domaines autres que l'édition scientifique, notamment dans des textes de communication informels ou à faible visibilité. Si une telle application dans l'édition scientifique n'est pas à proscrire catégoriquement, bien sûr en respectant les précautions nécessaires et dans des conditions bien encadrées, il faut se demander si un tel usage répond aux objectifs de valorisation et visibilité de la production scientifique francophone et multilingue en général.

[24] Cairn International Edition: Your gateway to the francophone social sciences and humanities

Une fonctionnalité de traduction automatique sans post-édition pourrait certes être utile pour des lecteurs savants, qui pourraient ainsi traduire automatiquement des publications potentiellement intéressantes afin d'en comprendre le sens général et éventuellement décider de creuser le travail des confrères. Ce type de lectorat serait par ailleurs capable de détecter et combler les failles de la traduction automatique grâce à des connaissances disciplinaires pointues. Bien différent est le discours pour le lectorat généraliste, qui pourrait consulter des contenus scientifiques par curiosité sans forcément avoir les connaissances nécessaires pour détecter des problèmes éventuels de traduction automatique. Néanmoins, dans un tel type d'usage sans finalités de recherche, la prise de risque en cas de mauvaise compréhension demeure limitée.

Au-delà de ces considérations, il est cependant fondamental de se poser les questions suivantes : ces traductions automatiques instantanées, qui *de facto* seraient des textes « éphémères » ne faisant pas l'objet d'un travail éditorial, peuvent-elles être citées ? Peuvent-elles être indexées et référencées systématiquement ? Sans doute non. Par ailleurs, il ne faudrait pas non plus favoriser la circulation de fausses informations pouvant découler de faux positifs de traduction automatique ; parfois, le texte brut traduit automatiquement est tellement fluide que les erreurs deviennent presque indétectables lors d'une simple lecture.

Celle de la traduction automatique sans post-édition n'est donc pas une piste à exclure *a priori*. Il sera en revanche nécessaire d'effectuer des évaluations rigoureuses et des choix raisonnés en fonction des disciplines et des types de publications, sans oublier les contraintes liées à la dette technique des différentes plateformes et infrastructures de publication. Il faudra par ailleurs s'assurer de pouvoir proposer des traductions automatiques de qualité acceptable pour encourager les utilisateurs à adopter la fonctionnalité. Dans tous les cas, une telle utilisation devra toujours être signalée et encadrée. Dans ce contexte, il convient de rappeler les principes d'utilisation responsable de la traduction automatique à des fins d'information, évoqués par M.J. Martindale[25] :

1. Aucune décision ne doit être prise sur la base d'une traduction automatique non vérifiée par un réviseur humain. Tout comme les traductions professionnelles, la traduction automatique doit faire l'objet d'un contrôle qualité.
2. Ne pas déployer ou encourager l'utilisation de moteurs de traduction automatique sans avoir préalablement effectué des évaluations indépendantes dans les domaines concernés. Aucune supposition sur la qualité d'un moteur doit être faite sur la base d'évaluations hors domaine ou de paires linguistiques non pertinentes.
3. Proposer des formations. Aider les utilisateurs à comprendre la technologie et les possibles variations en termes de qualité.
4. Montrer plusieurs propositions de traduction automatique sur le même écran. Les moteurs pourraient commettre des erreurs différentes et les utilisateurs pourraient mieux les identifier en remarquant ces différences.
5. Permettre un accès facile, voire automatisé, à des ressources linguistiques supplémentaires. Les dictionnaires, les glossaires et les bases terminologiques, les mémoires de traduction et d'autres outils souvent utilisés par les traducteurs pourraient également être mis à disposition des utilisateurs, à condition que ces outils soient suffisamment faciles à utiliser.

[25] M.J. Martindale, 2020, Responsible 'Gist' MT Use in the Age of Neural MT, Conférence annuelle AMTA 2020 - Association for Machine Translation in the Americas

6. Fournir des outils pour aider les utilisateurs à reconnaître les problèmes de qualité : évaluations et comparaisons automatiques, alignement entre le texte source et la traduction automatique, etc.

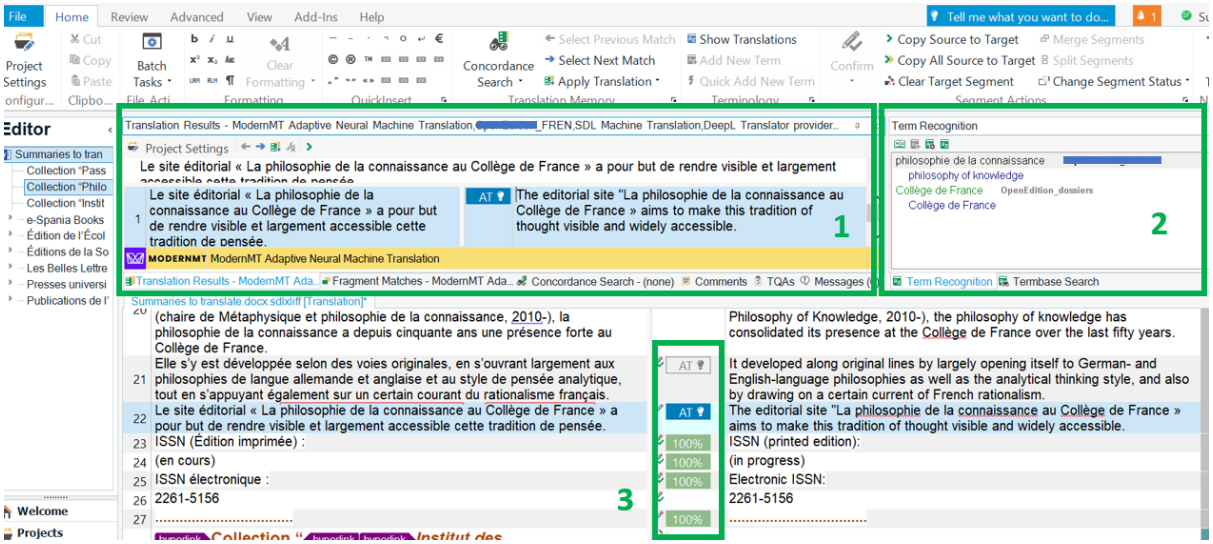
Pour conclure, aucune piste ne peut être écartée à ce stade. Il faudra mener davantage d'études sur les pratiques de traduction, connues ou possibles, ainsi que sur leur viabilité. Des solutions hybrides, impliquant donc plusieurs types d'acteurs (chercheurs-auteurs, traducteurs professionnels, étudiants supervisés, éditeurs, plateformes et grandes infrastructures), semblent être la voie à privilégier pour pouvoir garantir une systématisation et une durabilité du multilinguisme dans l'édition scientifique.

2.5 Inventaire d'outils de traduction automatique et assistée par ordinateur

Depuis l'avènement de la traduction automatique neuronale, les technologies de la traduction ont suscité un engouement surprenant auprès du grand public et d'acteurs spécialisés. Les développements et les progrès s'enchaînent, en rendant le paysage très riche et diversifié : outils libres ou commerciaux, génériques ou spécialisés, disponibles en ligne ou intégrés dans des environnements informatiques spécifiques, le choix est vraiment très vaste.
Dans le cadre d'une édition scientifique qui se veut ouverte et accessible, il apparaît important de proposer, à terme, des outils ouverts ou semi-ouverts, libres de logiques de profit et adaptés en termes d'ergonomie, accessibilité et performance. Conscient du temps nécessaire pour atteindre cet objectif ambitieux mais incontestable, le groupe de travail a élaboré un inventaire comparatif, n'ayant pas vocation à être exhaustif, afin d'identifier de possibles solutions d'attente.

L'inventaire est constitué de deux sections : une dédiée aux outils de traduction automatique et une consacrée aux outils de traduction assistée par ordinateur (TAO). Selon les meilleures pratiques professionnelles, la post-édition de traduction automatique s'effectue en effet dans le cadre de l'environnement technologique et des processus de travail proposés par les outils de TAO.

Exemple d'environnement de travail professionnel pour un projet de post-édition de traduction automatique[26]



[26] Dans la fenêtre 1, le traducteur retrouve les résultats issus des mémoires de traduction activées et des moteurs de traduction automatique associés par API. Les libellés signalés par le numéro 3 (pourcentages, AT, NMT) apparaissent à côté de chaque segment au moment de la traduction pour donner au traducteur une indication sur la provenance de la traduction proposée par l'outil de TAO. Si le libellé indique un pourcentage, la proposition est issue d'une mémoire de traduction et ce pourcentage indique l'analogie entre le segment traduit enregistré en mémoire et le segment en cours de traduction. Si le libellé indique les acronymes AT (*Automatic Translation*) ou NMT (*Neural Machine Translation*), la proposition est issue de l'un des moteurs de traduction automatique associés par API. Dans la fenêtre 2, le traducteur retrouve les consignes terminologiques selon le contenu des glossaires et des bases terminologiques activés. Les fonctionnalités de contrôle qualité de l'outil de TAO aident à repérer les incohérences éventuelles lors de la validation du segment post-édité par le traducteur.

Pour chaque outil listé dans l'inventaire, les aspects suivants sont analysés :

- **Accessibilité :** informations sur la nature et l'accessibilité de l'outil (commercial ou libre, accessible gratuitement, soumis à abonnement, etc.)
- **Ouverture du code source**
- **Ergonomie :** informations sur le processus permettant d'obtenir les traductions
- **Alimentation (uniquement pour les outils de traduction automatique) :** informations sur les processus d'alimentation des outils de traduction automatique (moteur générique pré-entraîné, moteur spécialisé nécessitant des corpus et/ou des informaticiens pour l'alimentation et le développement, processus d'alimentation simplifié pouvant être effectué en autonomie, etc.)
- **Cas d'utilisation :** informations sur les possibilités d'utilisation des outils (aide à la traduction et à la rédaction, possibilités d'intégration dans la chaîne éditoriale, etc.)
- **Public visé :** identification des publics qui pourraient utiliser les outils (chercheurs, traducteurs ou autres acteurs sollicités par les éditeurs ou les plateformes pour effectuer et gérer les traductions, lecteurs, etc.) selon les modalités d'accès et de fonctionnement de l'outil
- **Avantages :** ouverture ou semi-ouverture de l'outil, protection des données, qualité de sortie, adaptativité, interopérabilité, etc.
- **Inconvénients :** manque de contrôle (effet « boîte noire »), problèmes de qualité, discontinuité en cas de désabonnement, récupération et utilisation des données renseignées de la part du producteur du moteur, etc.

2.5.1 Outils de traduction automatique

Les outils sont listés par ordre alphabétique.

1. DeepL

Accessibilité	Moteur commercial pouvant être utilisé gratuitement
Ouverture du code source	Non
Ergonomie	Interface en ligne permettant d'obtenir la traduction : - en copiant le texte à traduire dans l'espace prévu à cet effet (mode "copier-coller") - en téléversant un fichier qui sera traduit dans son intégralité au même format (formats disponibles : Word, PowerPoint, texte .txt). Fonctionnalité soumise à limitations dans la version gratuite mais aussi dans les forfaits payants (de 5 à 100 par mois)
Alimentation	Moteur générique pré-entraîné avec possibilité de personnalisation terminologique pour certaines paires linguistiques (possibilité de créer un glossaire en saisissant le terme dans une langue et son équivalent dans l'autre - fonctionnalité actuellement disponible uniquement dans l'interface en ligne, pas dans les outils de TAO)

Cas d'utilisation	Moteur largement utilisé par la communauté scientifique grâce à sa facilité d'utilisation et à ses performances
Public visé	- Chercheurs - Traducteurs - Lecteurs
Avantages	- Ergonomie - Grande fluidité - Outil d'aide à la traduction et à la rédaction en langue étrangère, dont l'efficacité a été prouvée
Inconvénients	- Résultats parfois inconstants malgré les apparences - La grande fluidité peut cacher des erreurs graves qui deviennent presque indétectables - Les résultats varient selon le niveau de spécialisation et le genre textuel - Pas de cohérence des équivalences au niveau du document - Pas d'intégration aux outils de TAO - **Pas de protection des données**

2. DeepL Pro

Accessibilité	Moteur commercial soumis à abonnement (prix de l'abonnement à partir de 5,99 € par mois)
Ouverture du code source	Non
Ergonomie	Interface en ligne permettant d'obtenir la traduction : - en copiant le texte à traduire dans l'espace prévu à cet effet (mode "copier-coller") - en téléversant un fichier qui sera traduit dans son intégralité au même format (formats disponibles : Word, PowerPoint, texte .txt). Fonctionnalité soumise à limitations dans la version gratuite mais aussi les forfaits payants (de 5 à 100 par mois) Clé API permettant d'intégrer la traduction automatique dans les principaux outils de TAO (SDL Trados, memoQ, Déjà Vu, Wordfast Classic, Wordfast Anywhere e Wordfast Pro, Across, Memsource)
Alimentation	Moteur générique pré-entraîné avec possibilité de personnalisation terminologique pour certaines paires linguistiques (possibilité de créer un glossaire en saisissant le terme dans une langue et son équivalent dans l'autre - fonctionnalité actuellement disponible uniquement dans l'interface en ligne, pas dans les outils de TAO)
Cas d'utilisation	Moteur largement utilisé par la communauté scientifique grâce à sa facilité d'utilisation et à ses performances. Contrairement à la version gratuite, DeepL Pro permet une utilisation intégrée dans

	les outils de TAO, selon les standards les plus répandus dans le milieu de la traduction professionnelle
Public visé	- Chercheurs - Traducteurs - Lecteurs
Avantages	- Ergonomie - Grande fluidité - Outil d'aide à la traduction et à la rédaction en langue étrangère, dont l'efficacité a été prouvée - **Protection des données**
Inconvénients	- Résultats parfois inconstants malgré les apparences - La grande fluidité peut cacher des erreurs graves qui deviennent presque indétectables - Les résultats varient selon le niveau de spécialisation et le genre textuel - Pas de cohérence des équivalences au niveau du document - Discontinuité en cas de désabonnement

3. eTranslation

Accessibilité	Moteur public développé par la Commission européenne, pouvant être utilisé gratuitement sous certaines conditions : à l'heure actuelle, le service est destiné aux administrations publiques, aux petites et moyennes entreprises et aux départements de langues des universités européennes, ainsi qu'aux projets relevant du mécanisme pour l'interconnexion en Europe.
Ouverture du code source	Non
Ergonomie	Interface en ligne permettant d'obtenir la traduction : - en copiant le texte à traduire dans l'espace prévu à cet effet (mode "copier-coller") - en téléversant un fichier qui sera traduit dans son intégralité soit au même format (formats disponibles : Word, Excel, PowerPoint, PDF, LibreOffice, texte .txt, fichiers de type .xliff, fichiers .tmx, HTML) soit en format exploitable par les outils de TAO (fichiers de type .xliff, fichiers .tmx)
Alimentation	Moteur pré-entraîné spécialisé par domaine (notamment domaines institutionnels)
Cas d'utilisation	Moteur utilisé par les institutions et les administrations publiques, les PMEs et les universités européennes. Aucun usage connu dans l'édition scientifique mis à part une expérimentation dans le domaine médical.
Public visé	- Traducteurs

	- Chercheurs
Avantages	- Protection des données - Bonne qualité de sortie - Corpus conséquents et qualitatifs - Outil à gestion publique
Inconvénients	- Accès et déploiement à grande échelle actuellement difficile à cause des prérequis à remplir pour l'utilisation - Moteur non entraîné pour les sciences et l'édition scientifique en général (langage institutionnel) - Pas d'API disponible - Pas d'orientations vers l'open source pour l'instant

4. Google Traduction

Accessibilité	Moteur commercial pouvant être utilisé gratuitement
Ouverture du code source	Non
Ergonomie	Interface en ligne permettant d'obtenir la traduction en copiant le texte à traduire dans l'espace prévu à cet effet (mode "copier-coller")
Alimentation	Moteur générique pré-entraîné
Cas d'utilisation	Moteur largement utilisé par de nombreux publics pour sa notoriété, sa facilité d'utilisation et ses bonnes performances
Public visé	- Chercheurs - Traducteurs - Lecteurs
Avantages	- Ergonomie - Bonne qualité de sortie selon les paires linguistiques
Inconvénients	- Pas de protection des données - Résultats inconstants selon le niveau de spécialisation et le genre textuel - Nombre de caractères limité, exigeant de couper les documents en tranches

5. Inten.to

Accessibilité	API commerciale soumise à abonnement (prix sur devis)
Ouverture du code source	Des moteurs accessibles par l'API, certains sont basés sur un code open source, d'autres non
Ergonomie	API permettant d'obtenir des traductions automatiques issues de plusieurs moteurs en installant un plugin. Le plugin peut être

	installé sur : - Google Chrome - Microsoft Word, Excel et Outlook - Des sites intranet et outils spécifiques - Des outils de TAO (SDL Trados, MemoQ et XTM)
Alimentation	L'API donne accès à plusieurs moteurs, génériques ou personnalisables. Parmi ces moteurs, Amazon Translate, Google Cloud Translation, DeepL, Microsoft Translator, ModernMT et Systran (liste complète)
Cas d'utilisation	Grâce à sa souplesse, cette API permet de répondre à des besoins de traduction variés : traduction informative de documents internes ou peu visibles, traductions à haute visibilité avec post-édition, traductions dans différents domaines et services nécessitant des moteurs spécifiques, etc.
Public visé	- Chercheurs - Traducteurs - Lecteurs
Avantages	- Grande souplesse : l'API peut être utilisée dans plusieurs interfaces - Accès à plusieurs moteurs : cela permet de choisir son moteur selon le type de texte et de besoin à l'aide d'un système de *smart routing* - Protection des données - Possibilité de personnalisation
Inconvénients	- Efforts et coûts de déploiement importants dans le cadre d'un projet à grande échelle dans l'édition scientifique - Pas d'usages connus par ce groupe de travail dans l'édition scientifique - Les moteurs disponibles ne sont pas forcément les plus adaptés à l'édition scientifique - Discontinuité en cas de désabonnement

6. KantanMT

Accessibilité	Moteur commercial soumis à abonnement (prix sur devis)
Ouverture du code source	Oui
Ergonomie	- Clé API permettant d'intégrer la traduction automatique dans des outils de TAO (MemoQ, SDL Trados Studio, MemSource, XTM, Across and the Okapi Framework) - Plugin permettant d'intégrer la traduction automatique dans des sites web, des navigateurs (Google Chrome et Mozilla Firefox), la suite Microsoft Office et plusieurs types de documents.
Alimentation	Moteur personnalisable avec des données mises à disposition par

	le producteur ou des données propriétaires
Cas d'utilisation	Moteur utilisé par des entreprises et dans le milieu universitaire anglo-saxon, et notamment irlandais
Public visé	- Chercheurs - Traducteurs - Lecteurs
Avantages	- Possibilité de personnalisation - Processus de personnalisation rapide et facile, pouvant être effectué en autonomie par des profils non-ingénieurs - Protection des données - Moteur basé sur une technologie open source
Inconvénients	- Pas d'usages connus par ce groupe de travail dans l'édition scientifique - Discontinuité en cas de désabonnement

7. ModernMT

Accessibilité	Moteur commercial soumis à abonnement (prix de l'abonnement 50 $ par million de caractères)
Ouverture du code source	Oui
Ergonomie	- Interface en ligne permettant d'obtenir la traduction en copiant le texte à traduire dans l'espace prévu à cet effet (mode "copier-coller") - Clé API permettant d'intégrer la traduction automatique dans des outils de TAO (MateCat, SDL Trados, MemoQ)
Alimentation	Moteur pré-entraîné et adaptatif permettant un apprentissage en temps réel lors de l'activité de post-édition ou en important en quelques clics une mémoire de traduction
Cas d'utilisation	Moteur de plus en plus utilisé par des entreprises et des traducteurs pour sa simplicité d'utilisation, déploiement et personnalisation
Public visé	- Chercheurs - Traducteurs - Lecteurs
Avantages	- Moteur basé sur une technologie open source - Bonne qualité de sortie avec possibilité de personnalisation - Processus de personnalisation rapide et facile, pouvant être effectué en autonomie par des profils non-ingénieurs - Protection des données
Inconvénients	- Pas d'usages connus par ce groupe de travail dans

	l'édition scientifique - L'apprentissage en temps réel n'est pas toujours efficace - Sortie moins fluide que d'autres solutions similaires - Pas toujours cohérent au niveau du document - Discontinuité en cas de désabonnement

8. SDL Machine Translation

Accessibilité	Moteur commercial soumis à abonnement (prix sur devis)
Ouverture du code source	Non
Ergonomie	Moteur accessible depuis l'interface de l'outil de TAO SDL Trados
Alimentation	Moteur générique pré-entraîné
Cas d'utilisation	Moteur largement utilisé par les traducteurs car il est mis à disposition gratuitement avec la licence SDL Trados
Public visé	- Traducteurs
Avantages	- Bonne qualité de sortie - Protection des données
Inconvénients	- Utilisation du moteur liée à l'outil de TAO SDL Trados - Pas de possibilité de personnalisation - Discontinuité en cas de désabonnement

9. Systran Translate Pro

Accessibilité	Moteur commercial soumis à abonnement (prix sur devis)
Ouverture du code source	Oui
Ergonomie	Moteur permettant de traduire des textes, des documents entiers (MS Word, PowerPoint, PDF) et des pages web via son interface. Par clé API, il est possible d'intégrer la traduction automatique dans les principaux outils de TAO (SDL Trados, memoQ, Déjà Vu, Wordfast, Across, Memsource, Smartcat, Smartling, XTM)
Alimentation	Moteur générique pré-entraîné avec possibilité de personnalisation (corpus et terminologie)
Cas d'utilisation	Moteur utilisé par des entreprises du secteur privé dans des domaines comme la sécurité, les technologies et l'e-commerce, ainsi que par des institutions publiques. Des usages ont été identifiés également dans l'édition scientifique (plateformes et éditeurs)
Public visé	- Chercheurs

	- Traducteurs - Lecteurs
Avantages	- Protection des données - Bonne qualité de sortie avec possibilité de personnalisation (dictionnaire terminologique personnel ou choix de dictionnaire : science, life science, colloquial dictionary, etc.) - Basé sur la technologie open source OpenNMT - *Marketplace* permettant d'échanger des modèles avec d'autres acteurs pertinents
Inconvénients	- L'introduction d'un dictionnaire personnalisé génère des erreurs de déterminants - Terminologie pas toujours cohérente au sein du document - Discontinuité en cas de désabonnement

2.5.2 Outils de traduction assistée par ordinateur

Les outils sont listés par ordre alphabétique.

1. MateCat

Accessibilité	Outil de TAO gratuit
Ouverture du code source	Oui
Ergonomie	Interface intuitive accessible depuis les navigateurs web Google Chrome et Safari. Possibilité d'intégration dans des systèmes plus complexes.
Cas d'utilisation	Utilisé par des traducteurs professionnels pour ses fonctionnalités de TAO (mémoires de traduction, glossaires, contrôle qualité...) et d'intégration de la traduction automatique (compatible avec les moteurs ModernMT, AltLang, Apertium, Google Translate, Inten.to, Iconic, Microsoft Translator, Moses, Tilde, Yandex)
Public visé	- Chercheurs - Traducteurs
Avantages	- Interface intuitive - Gère de nombreux formats - Sans installation - Open source - Système de sauvegarde temporaire en cas de coupure d'Internet
Inconvénients	- Inaccessible depuis Mozilla FireFox et Internet Explorer

2. OmegaT

Accessibilité	Outil de TAO gratuit
Ouverture du code source	Oui
Ergonomie	Logiciel nécessitant une installation sur l'ordinateur
Cas d'utilisation	Utilisé par des traducteurs professionnels pour ses fonctionnalités de TAO (mémoires de traduction, glossaires, contrôle qualité...) et d'intégration de la traduction automatique (compatible avec les moteurs DeepL, Moses, Apertium, Google Translate, IBM Watson, Microsoft Translator, Yandex)
Public visé	- Chercheurs - Traducteurs
Avantages	- Open source
Inconvénients	- Nécessite une installation sur chaque poste utilisateur - Légèrement moins intuitif que ses homologues en ligne

3. SDL Trados

Accessibilité	Outil de TAO commercial soumis à l'achat d'une licence (prix selon utilisation)
Ouverture du code source	Non
Ergonomie	Logiciel nécessitant une installation sur l'ordinateur
Cas d'utilisation	Outil le plus répandu dans le milieu de la traduction professionnelle pour ses fonctionnalités de TAO (mémoires de traduction, glossaires, contrôle qualité...) et d'intégration de la traduction automatique (compatible avec les principaux moteurs de traduction automatique). Usages connus dans l'édition scientifique
Public visé	- Traducteurs
Avantages	- Fonctionnalités avancées et performantes - Très répandu parmi les traducteurs professionnels - Gère de nombreux formats
Inconvénients	- Compatible uniquement avec le système d'exploitation Windows - L'outil n'est pas open source - Nécessite une licence payante et une installation sur chaque poste utilisateur

4. Smartcat

Accessibilité	Outil de TAO gratuit
Ouverture du code source	Non
Ergonomie	Interface intuitive accessible depuis les principaux navigateurs web
Cas d'utilisation	Utilisé par des traducteurs professionnels pour ses fonctionnalités de TAO (mémoires de traduction, glossaires, contrôle qualité...) et d'intégration de la traduction automatique (Google Traduction, Amazon Translate, Baidu Translate, DeepL, Yandex, Microsoft Translator…)
Public visé	- Chercheurs - Traducteurs
Avantages	- Interface intuitive - Gère de nombreux formats - Sans installation - Système de sauvegarde temporaire en cas de coupure d'Internet
Inconvénients	- L'outil n'est pas open source

5. Wordfast Anywhere

Accessibilité	Outil de TAO gratuit
Ouverture du code source	Non
Ergonomie	Interface intuitive accessible depuis les principaux navigateurs web
Cas d'utilisation	Utilisé par des traducteurs professionnels pour ses fonctionnalités de TAO (mémoires de traduction, glossaires, contrôle qualité...) et d'intégration de la traduction automatique (Google Traduction, DeepL, Yandex, Microsoft Translator, Systran, KantanMT, Inten.to…)
Public visé	- Chercheurs - Traducteurs
Avantages	- Interface intuitive - Gère de nombreux formats de fichiers - Sans installation
Inconvénients	- L'outil n'est pas open source

2.5.3 Conclusions

Le paysage des technologies de la traduction est de plus en plus riche. Le présent inventaire basé sur les expériences des membres du groupe de travail n'offre qu'un aperçu de cette richesse mais il devrait permettre d'envisager des solutions pouvant favoriser le multilinguisme dans l'édition scientifique. Il faut néanmoins noter que si ces outils ont aujourd'hui fait leurs preuves dans le milieu de la traduction professionnelle dans différents domaines, leur introduction et déploiement à grande échelle dans l'édition scientifique reste un défi. Pour imaginer des scénarios de travail concrets, il faudra clairement identifier les acteurs qui interviendront dans les processus afin de tenir compte de leurs profils, compétences et attentes.

2.6 Constitution de bases de test et d'apprentissage

Afin d'évaluer la qualité des moteurs identifiés et de déployer à terme des moteurs ouverts ou semi-ouverts entraînés avec des données spécialisées, il sera nécessaire de constituer des corpus de textes disciplinaires qui serviront de bases de test et d'apprentissage. Ces deux types de bases seront utilisés respectivement pour évaluer la qualité de la traduction automatique et pour alimenter les moteurs afin d'en améliorer les résultats. La distinction entre les jeux de données est en effet nécessaire pour ne pas fausser l'évaluation ; en entraînant le moteur avec les données de test, celui-ci saura forcément produire de meilleurs résultats.
Les corpus seront donc utilisés pour extraire des données terminologiques et linguistiques spécifiques aux différents domaines, évaluer et alimenter des moteurs personnalisables afin que les textes traduits automatiquement soient les plus pertinents possibles d'un point de vue disciplinaire (pour plus d'informations sur l'évaluation de moteurs de traduction automatique, consulter la section 2.7 Principes d'évaluation et de post-édition de traduction automatique).

Aujourd'hui les corpus spécialisés adaptés à l'entraînement de moteurs de traduction automatique sont une ressource très recherchée. Les *marketplaces* et les offres commerciales se multiplient pour répondre à cette demande croissante.
L'objectif de déploiement à grande échelle de la traduction automatique dans l'édition scientifique étant assez inédit, il sera nécessaire de constituer des corpus disciplinaires *ad hoc* avec des textes monolingues, multilingues et bilingues, parallèles comme comparables. Pour rappel, un corpus comparable « contient deux (ou plus) sous-corpus qui sont indépendants l'un de l'autre : il peut s'agir du même type de texte dans deux langues originales différentes » ; au contraire, un corpus parallèle « est constitué de textes originaux en langue A et de leur traduction en langue B. »[27]

La collecte du projet Traductions et science ouverte portera sur les textes intégraux des publications ainsi que sur leurs métadonnées, en particulier les titres, les résumés et les mots-clés. Plusieurs sources ont déjà été identifiées parmi les éditeurs, les plateformes de contenus, les infrastructures et les réseaux savants et universitaires. Cependant, des défis se posent notamment en matière de conditions d'utilisation et de formats d'exploitation des données. Par exemple, les universités disposent parfois de corpus disciplinaires constitués dans le cadre de leurs formations mais souvent ces données ne peuvent pas être diffusées

[27] Rudy Loock, 2016, La traductologie de corpus. Villeneuve D'Ascq, Presses universitaires du Septentrion, 261 p.

car certains intervenants ne donnent pas les autorisations nécessaires. Ou encore, concernant les éditeurs et les plateformes de contenus, les différences dans les formats de mise à disposition des corpus pourraient compliquer encore davantage le travail de traitement et d'harmonisation, déjà complexe en soi.
Par ailleurs, la plupart des acteurs consultés pourront fournir surtout des corpus monolingues ou comparables avec très peu de corpus parallèles, pourtant fondamentaux pour l'apprentissage. Une solution pourrait être l'utilisation de la traduction automatique brute pour la constitution de corpus parallèles artificiels, pratique identifiée comme rétro-traduction, ou *back-translation*, dans le jargon technique, dont l'efficacité a été mise en lumière plusieurs fois[28]. Mais tout cela montre qu'un travail conséquent de mise en forme de données, sans doute spécifique à chaque fournisseur de corpus, sera nécessaire ; cette étape, préalable à toute autre action, vise donc également à jeter les bases d'une interopérabilité des données, même s'il est compliqué aujourd'hui d'estimer avec précision la durée et le coût du chantier.

Une autre question qui se pose concerne l'ouverture des bases de test et d'apprentissage. Une ouverture des données serait souhaitable pour favoriser une dynamique de recherche plus large autour de la traduction automatique dans l'édition scientifique. Or, les acteurs qui ont accepté de mettre à disposition des corpus le font sous des conditions juridiques bien précises qui ne permettent pas d'envisager une mise à disposition plus large, par exemple à d'autres acteurs R&D qui pourraient contribuer au développement et à l'amélioration de solutions personnalisées. La directive européenne sur le droit d'auteur et les droits voisins dans le marché unique numérique, adoptée en 2019 et actuellement en cours de transposition en droit français, autorise la fouille de texte et de données, ce qui aura un impact sur le développement de la traduction automatique et d'autres activités d'analyse de la production scientifique : comparaison d'articles, détection de similarités, voire de plagiats, élaboration de cartographies, etc. Cependant, il ne sera pas possible de mettre à disposition de partenaires les corpus et les bases de test et d'apprentissage élaborées pour favoriser les dynamiques collaboratives et le développement d'outils fondés sur l'exploitation de ces bases.
Un autre argument en faveur de l'ouverture des bases de test et d'apprentissage serait la possibilité de créer des réseaux de coopération internationale pour favoriser la constitution de corpus disciplinaires multilingues, avec des financements mis à disposition par la Commission européenne et par d'autres initiatives et réseaux spécialisés : European Language Grid, EAMT, ELRC, ELRA... Une telle démarche permettrait de gagner en visibilité, en qualité, ainsi qu'en moyens et ressources pouvant être plus facilement mutualisées et diffusées pour des fins de recherche.

2.7 Principes d'évaluation et de post-édition de traduction automatique

Comme déjà mentionné, l'offre de moteurs de traduction automatique est aujourd'hui très riche et la qualité des résultats peut sensiblement varier selon les domaines, les contextes d'utilisation, les combinaisons linguistiques et d'autres facteurs. Avant d'adopter un moteur de traduction automatique dans un contexte donné, il est donc nécessaire d'effectuer une évaluation sur la base de critères spécifiques et rigoureux. De la même manière, les consignes

[28] Patrik Lambert, 2020, Unsupervised Adaptation of Neural MT with Iterative Back-Translation, Iconic blog issue No.67 [consulté le 19/11/2020]

de post-édition fournies aux traducteurs ou à d'autres intervenants appelés à améliorer les résultats de la traduction automatique peuvent varier selon les objectifs d'utilisation et les attentes de qualité pour un contenu donné. Cette section du rapport vise à fournir des orientations générales qui pourront être adaptées aux spécificités des projets menés à l'avenir. De manière générale, l'objectif est d'évaluer et sélectionner un ou plusieurs moteurs de traduction automatique, dont la sortie sera soumise à un travail de post-édition afin d'obtenir des traductions de qualité publiable d'un point de vue linguistique et éditorial.

2.7.1 Principes d'évaluation de la traduction automatique

Comment évalue-t-on la qualité d'un moteur de traduction automatique ? La question est loin d'être anodine, car plusieurs facteurs entrent en jeu lors de l'évaluation d'une traduction, qu'elle soit humaine ou automatique.

D'abord, il faut considérer que la traduction n'est pas une science exacte : un même texte source peut avoir plusieurs traductions, toutes également acceptables. Si l'évaluation d'une traduction est confiée à des évaluateurs humains, la notion de subjectivité intervient également dans le processus d'évaluation ; il n'est pas rare, en effet, que les évaluateurs ne soient pas d'accord sur le niveau de qualité d'une même traduction. Afin de pallier les biais de la subjectivité, il est donc primordial de bien définir les objectifs et les indicateurs de chaque évaluation. La subjectivité n'est pas le seul inconvénient de l'évaluation humaine ; il s'agit également d'un processus coûteux en temps et en ressources. En alternative à l'évaluation humaine, il est donc possible de recourir à des algorithmes pour effectuer une évaluation automatique[29] qui, certes, a le mérite d'être moins chère et plus rapide par rapport à l'évaluation humaine, mais est également moins précise. Les deux types d'évaluation présentent donc des avantages et des inconvénients ; encore une fois, il convient de prendre en compte tous les besoins afin de faire un choix avisé.

De manière générale, le groupe de travail suggère d'adopter un processus d'évaluation humaine avec des traductions dites « de référence ». Cette méthode, qui peut bien évidemment être complétée par des métriques automatiques, consiste à fournir à un évaluateur humain un texte source, une ou plusieurs traductions humaines et une ou plusieurs traductions automatiques. L'évaluateur comparera les différentes traductions automatiques avec le texte source et les traductions humaines de référence, afin d'établir un classement sur la base de critères objectifs.

Dans le cadre de l'édition scientifique, il convient néanmoins de souligner que les traductions humaines actuellement publiées, qui serviront donc de traductions de référence, peuvent avoir des niveaux de qualité très hétérogènes. En effet, comme démontré par un article récemment publié[30], l'évaluation d'une traduction automatique par comparaison avec des traductions humaines n'est complètement fiable que si ces dernières sont de bonne qualité : elles ne doivent donc pas comporter d'erreurs de compréhension ou de fluidité, ni de maladresses résultant d'un travail de post-édition incomplet, et elles ne doivent pas être surcorrigées pour améliorer la fluidité au détriment de la précision. Les évaluateurs sollicités dans le cadre du

[29] Automated MT Evaluation Metrics, TAUS [consulté le 19/11/2020]

[30] S. Läubli, S. Castilho, G. Neubig, R. Sennrich, Q. Shen, A. Toral, 2020, A Set of Recommendations for Assessing Human-Machine Parity in Language Translation, Journal of Artificial Intelligence Research 67, 653-672

projet Traductions et science ouverte devront donc tenir compte de la qualité effective des traductions humaines publiées pour ne pas fausser les résultats de l'évaluation. Dans ce sens, le groupe propose d'identifier ces traductions publiées comme « traductions de départ » au lieu de « traductions de référence ».

L'évaluation des moteurs de traduction automatique sera donc effectuée en amont de tout projet opérationnel sur une sélection de textes intégraux et métadonnées extraits des corpus parallèles récupérés (voir section 2.6 Constitution de bases de test et d'apprentissage). Pour l'évaluation seront sollicités des traducteurs natifs de la langue cible et spécialisés, ainsi que des experts dans les disciplines sur lesquels porteront les tests.

La grille d'évaluation communiquée aux évaluateurs sera construite à partir de celle élaborée par les chercheurs de l'axe TRASILT de l'équipe LIDILE[31]. Disponible sous licence Creative Commons, cette grille repose sur neuf catégories d'erreurs, dont sept catégories plutôt traditionnelles (Sens, Grammaire / syntaxe, Orthographe / typographie, Terminologie, Phraséologie, Style, Omission / ajout) et deux catégories issues de l'évaluation professionnelle : les erreurs de Localisation (définies comme l'absence d'adaptation à un public cible ou à une culture donnée) et les erreurs de PAO (défauts de mise en page et de formatage). À chacune des erreurs identifiées, peuvent s'appliquer potentiellement quatre effets sur la qualité de la traduction évaluée : la Précision de l'information transmise, la Fonctionnalité du document traduit, la Lisibilité des contenus et la Conformité de la traduction aux différentes normes et conventions linguistiques ou professionnelles applicables. Ces indicateurs seront complétés par une mesure partagée de l'effort de post-édition et par des indicateurs permettant, par exemple, d'évaluer l'ergonomie et l'impact de la traduction automatique sur les processus de publication ou de distribution, avec ou sans post-édition.

Si l'évaluation est une étape initiale incontournable pour choisir le ou les moteurs de traduction automatique les plus pertinents pour chaque projet opérationnel, il conviendra de répéter régulièrement le processus évaluatif afin de s'assurer de la pertinence de l'usage de la traduction automatique dans les différents types de textes et pour mesurer les progrès en termes de qualité des résultats grâce à l'entraînement continu des moteurs.

2.7.2 Principes de post-édition de la traduction automatique

L'activité de post-édition consiste à réviser et corriger le texte brut pré-traduit automatiquement par un moteur afin d'atteindre le niveau de qualité souhaité. Ce niveau de qualité peut en effet varier selon les attentes et les usages prévus pour le contenu à traduire. Les consignes de post-édition seront donc adaptées en conséquence.

Dans le cadre du projet Traductions et science ouverte, le groupe de travail préconise d'atteindre par l'activité de post-édition des contenus pré-traduits automatiquement un niveau de qualité publiable d'un point de vue linguistique et éditorial. Si les consignes de post-édition devront être adaptées en fonction du flux de travail établi, et notamment des acteurs sollicités pour réviser la traduction automatique, il est d'ores et déjà possible de fournir des principes

[31] D. Toudic, K. Hernandez Morin, F. Moreau, F. Barbin, G. Phuez, 2014, Du contexte didactique aux pratiques professionnelles : proposition d'une grille multicritères pour l'évaluation de la qualité en traduction spécialisée, ILCEA [En ligne], 19 | 2014, consulté le 22 octobre 2020. URL : http://journals.openedition.org/ilcea/2517 ; DOI : https://doi.org/10.4000/ilcea.2517

généraux inspirés aux recommandations publiées par le réseau TAUS sous la direction de Sharon O'Brien (Dublin City University) et Fred Hollowood (Trinity College Dublin)[32] :

- Viser une traduction correcte au niveau grammatical, sémantique et syntaxique ;
- Vérifier qu'aucune information n'a été ajoutée ou omise dans la traduction ;
- Veiller à ce que la terminologie spécialisée soit correctement traduite ;
- Exploiter au maximum le résultat brut de la traduction automatique sans apporter des corrections stylistiques préférentielles ;
- Prêter une attention particulière aux noms propres, toponymes, acronymes, abréviations et d'autres éléments susceptibles de poser problèmes aux moteurs de traduction automatique ;
- Garder une trace du texte source, du texte pré-traduit automatiquement, des actions du post-éditeur et du résultat final ;
- Garder des informations précises des acteurs de la post-édition.

3. De la théorie à la pratique

3.1 Appel à projets Traductions scientifiques

Afin de répondre à la stagnation du référencement international de la recherche française et renforcer ainsi sa visibilité en dehors de la communauté francophone, le comité de suivi de l'édition scientifique (CSES) a lancé mi-décembre 2018 l'appel à projet Traductions scientifiques avec l'objectif de soutenir des expérimentations innovantes d'amélioration de la qualité des métadonnées multilingues. Portant une attention particulière à l'activité de traduction et notamment aux processus impliquant l'utilisation de solutions de traduction assistée et automatique, le jury a retenu les huit projets suivants :

- **Projet MetaPHora** « Métadonnées Plurilingues et Homogènes pour un collectif de Revues françaises en Archéologie »
  Discipline : archéologie
  Langues : anglais, allemand, arabe
  Revues partenaires : Archéologie médiévale, Gallia - Archéologie des Gaules, Gallia préhistoire, Préhistoire méditerranéenne

- **Projet CybergéoNet**
  Discipline : géographie et sciences sociales voisines
  Langues : anglais, espagnol, chinois
  Revues partenaires : Cybergeo, Brussels Studies, Via Tourism Review, Ar@cne, encyclopédie Hypergéo

- **Projet de traduction assistée par ordinateur d'articles médicaux des éditions John Libbey**
  Discipline : médecine
  Langues : anglais

[32] TAUS MT Post-editing Guidelines [consulté le 23 octobre 2020]

Revues partenaires : Gériatrie et Psychologie Neuropsychiatrie du Vieillissement, Hématologie

- **Projet Open Traduction**
  Discipline : orient, monde arabe
  Langues : anglais, arabe
  Revues partenaires : Bulletin de l'IFAO, Annales Islamologiques, Mélanges Institut dominicain études orientales, L'Année du Maghreb, Revue des mondes musulmans et de la Méditerranée, Arabian Humanities, Revue internationale d'archéologie et de sciences sociales sur la péninsule Arabique, Égypte/monde arabe, Insaniyat, Bulletin des Études orientales

- **Projet TesTradSHS** « Tester la traduction automatique dans les revues de Sciences humaines et sociales »
  Discipline : sciences humaines et sociales
  Langues : anglais
  Revues partenaires : Annales, Histoire, Sciences Sociales, Archives des sciences sociales des religions, Études rurales, L'Homme, European Journal of Turkish Studies

- **Projet « Amélioration des métadonnées de revues scientifiques »**
  Discipline : sciences politiques et sociales
  Langues : anglais
  Revues partenaires : Histoire@politique, Revue de l'OFCE, Raisons politiques, Revue française de science politique, 20 & 21. Revue d'histoire

- **Projet « Un outil de traduction automatique intégré à la chaîne éditoriale »**
  Discipline : sciences humaines et sociales interdisciplinaires ou pluridisciplinaires
  Langues : anglais
  Revues partenaires : Archéosciences, Norois, Annales de Bretagne, Éducation et Didactique

- **Projet TRanslens « Traductions, Diffusion, Réflexions, Expérimentations »**
  Discipline : sciences humaines et sociales interdisciplinaires ou pluridisciplinaires
  Langues : anglais, allemand, portugais, espagnol, polonais, et russe ou italien
  Revues partenaires : Biens symboliques - Symbolic Goods

Malgré les ralentissements liés à la crise sanitaire en 2020, le bilan provisoire des projets, dont la finalisation est prévue pour fin 2021, donne d'ores et déjà des indications précieuses pour les chantiers à venir, même s'il sera nécessaire d'attendre pour pouvoir tirer des conclusions plus exhaustives.
Le focus des projets sur les métadonnées des publications (titres, description bibliographique, mots-clés, résumé court ou long) montre qu'à l'heure actuelle ce besoin de traduction, pourtant fondamental pour la découvrabilité de la production scientifique française et internationale, non seulement n'est pas systématiquement satisfait, mais présente également une marge d'amélioration importante en termes de processus et de qualité des traductions. Par exemple, l'utilisation de ressources et de technologies mutualisées sur base disciplinaire pourrait être particulièrement pertinente pour la traduction des métadonnées ; il faudra donc étudier des solutions et des processus pour favoriser cette optimisation fondamentale.

La création et l'amélioration de ressources partagées, interopérables et accessibles aux intervenants dans le processus de traduction, apparaît comme prioritaire pour plusieurs porteurs de projets, qui ont indiqué vouloir créer des thésaurus et des mémoires de traduction mutualisés. Par ailleurs, les équipes des projets lauréats se sont naturellement tournées les unes vers les autres pour échanger sur les processus, les bonnes pratiques, les ressources et les résultats. Cela montre qu'une collaboration à plus grande échelle, en regroupant plusieurs éditeurs par discipline, est non seulement possible, mais également fructueuse.
À noter également que les projets emploient essentiellement des technologies propriétaires incompatibles avec des mécanismes d'ouverture et de personnalisation des moteurs. Si ces choix sont tout à fait compréhensibles compte tenu du temps et des ressources à disposition des équipes lauréates, il convient de souligner encore une fois l'importance d'utiliser des technologies ouvertes ou semi-ouvertes, en tout cas personnalisables, qui pourront être mises en place et entretenues en mutualisant les ressources.
Une fois ces technologies identifiées, testées et déployées, il faudra prévoir des formations à destination des différents usagers. Une difficulté rencontrée par plusieurs équipes a été en effet le recrutement de traducteurs spécialisés et cette difficulté a été d'autant plus importante quand le processus impliquait l'utilisation de technologies de la traduction (outils de traduction automatique mais aussi de TAO). Certains traducteurs n'ont pas l'habitude de les utiliser, d'autres les refusent catégoriquement. Cela confirme qu'un travail de sensibilisation et de formation sera nécessaire pour tous les intervenants dans le processus. Les technologies de la traduction ne font en effet preuve d'une réelle utilité que si elles sont maîtrisées par les intervenants et intégrées dans un environnement de travail complété par des ressources linguistiques fiables. Si ces ressources sont incomplètes ou inaccessibles, et si les intervenants ne sont pas à l'aise avec ces outils, leur utilisation génère de la frustration et une mauvaise réputation de ces technologies auprès des utilisateurs.

3.2 Actions et expérimentations recommandées

Dans la continuité de l'appel à projets Traductions scientifiques, le groupe de travail recommande à court et moyen terme les actions et les expérimentations listées ci-après par étapes de 1 à 7. À noter qu'il sera important de faire preuve de prudence et de dédier à chaque étape les temps de réflexion et de réalisation nécessaires afin que les solutions proposées répondent aux besoins spécifiques et soient acceptées par les membres des communautés concernées. L'utilisation des technologies de la traduction, et notamment de la traduction automatique, soulève en effet de nombreux questionnements techniques, opérationnels, éthiques et intellectuels qui ne doivent pas être négligés dans le cadre d'un déploiement à grande échelle dans un contexte si particulier et inédit comme l'édition scientifique. Selon les premières estimations, le chantier devrait s'étaler sur deux ans minimum (2021/2022).

1. Analyse de la nature et de la volumétrie des corpus multilingues identifiés et étude d'autres possibilités pour la collecte
2. Traitement des corpus collectés afin d'obtenir des bases de test et d'apprentissage et des ressources linguistiques mutualisées
3. Évaluation de moteurs de traduction automatique en utilisant les bases de test et d'apprentissage
4. Organisation de journées d'études rassemblant les porteurs des projets lauréats de l'appel Traductions scientifiques et d'autres acteurs pertinents

5. Création d'un démonstrateur pour préfigurer un processus de traduction à grande échelle
6. Élaboration d'un guide à destination des chercheurs et des institutions de recherche sur la traduction automatique, la rédaction en langue étrangère et la « rédaction claire » (adaptée à la traduction automatique)
7. Étude de pistes de collaboration dans les réseaux d'éditeurs européens pour la constitution de corpus

3.2.1 Analyse de la nature et de la volumétrie des corpus multilingues identifiés et étude d'autres possibilités pour la collecte

Afin d'avoir une idée plus précise de la nature et de la volumétrie des corpus identifiés et de repérer d'autres sources de corpus potentielles, il est souhaitable de mener une étude approfondie permettant notamment de préciser :

- La volumétrie et la nature éditoriale des corpus multilingues comparables chez les fournisseurs identifiés
- La volumétrie et la nature éditoriale des corpus multilingues parallèles chez les fournisseurs identifiés
- L'éventuelle existence de corpus pertinents chez d'autres fournisseurs

Ce travail, qui pourrait être confié à des profils de documentalistes, permettra de mieux recenser les ressources disponibles et planifier le traitement des données.

À l'heure actuelle, les pistes ci-dessous ont été identifiées pour la collecte des corpus disciplinaires (en ordre alphabétique) :
Cairn.info, CISMeF, Cochrane, Commission européenne, Éditions John Libbey, Éditions Quæ, Elsevier OA CC-By Corpus, Episciences, Érudit, HAL, Huma-Num, Inist-CNRS, Istex, MeSH bilingue, National Library of Medicine / Inserm, OCDE, OpenEdition, Opus, OSCAR, Scientext, Thésaurus Pactols, TAUS, theses.fr, Triple, UMLS, US Public Library of Science, WMT.

3.2.2 Traitement des corpus collectés afin d'obtenir des bases de test et d'apprentissage exploitables et des ressources linguistiques mutualisées

Le traitement des corpus collectés permettra d'obtenir des données exploitables dans le cadre des opérations de test et d'apprentissage des moteurs de traduction automatique ainsi que d'effectuer des extractions terminologiques pour la constitution de glossaires et de bases terminologiques multilingues. Ce travail apparaît comme prioritaire dans la planification des actions suggérées. Afin de pouvoir estimer la temporalité et les ressources nécessaires, le groupe de travail a demandé aux fournisseurs de corpus de recevoir des informations sur la volumétrie des données ainsi que des échantillonnages techniques, variés du point de vue éditorial, pour déterminer des actions et des traitements automatiques à mener en fonction du format et de la structuration des données : comme déjà évoqué, le langage XML se décline et s'instancie de manière variée chez les éditeurs.

3.2.3 Évaluation de moteurs de traduction automatique en utilisant les bases de test et d'apprentissage

Après avoir transformé les corpus collectés en bases de test et d'apprentissage, il sera possible de démarrer des évaluations qui permettront de sélectionner le ou les moteurs de traduction automatique les plus pertinents pour les différentes disciplines et combinaisons linguistiques (pour rappel, les tests démarreront sur la paire français-anglais dans les disciplines indiquées dans la section 2.1 Périmètre disciplinaire). Pour plus d'informations sur le processus d'évaluation, consulter la section 2.7.1 Principes d'évaluation de la traduction automatique.

3.2.4 Organisation de journées d'études rassemblant les porteurs des projets lauréats de l'appel Traductions scientifiques et d'autres acteurs pertinents

Afin de définir des prototypes de processus pouvant être déployés et mutualisés par plusieurs acteurs, il apparaît important d'attendre la finalisation des projets lauréats de l'appel Traductions scientifiques et de réunir ensuite les acteurs pertinents (éditeurs, plateformes de diffusion, chercheurs, enseignants et professionnels de la traduction, etc.) pour faire des propositions sur la base de constats et résultats concrets. Cela pourrait se faire dans le cadre de journées d'études dédiées à l'automne 2021. Ces journées seront l'occasion pour discuter des flux de travail à mettre en place, des intervenants à mobiliser et des compétences requises dans les différentes étapes du processus (post-édition de traduction automatique sur des textes scientifiques, validation des traductions avant publication, intégration dans la chaîne éditoriale, etc.)

3.2.5 Création d'un démonstrateur pour préfigurer un processus de traduction à grande échelle

Aujourd'hui, malgré leur intérêt pour l'enjeu des traductions, tous les éditeurs et les plateformes de contenus consultés n'ont pas nécessairement la possibilité de proposer des modules dédiés. Il serait donc souhaitable de développer une plateforme indépendante, initialement sous forme de preuve de concept, pour définir et tester des flux de traduction avec l'objectif d'un déploiement à grande échelle.

Ce démonstrateur permettrait aux éditeurs et aux plateformes de contenus de déposer des textes à traduire. Ces contenus seraient ensuite automatiquement transférés vers un outil de TAO associé ou intégré au démonstrateur pour permettre le travail de post-édition avec des ressources partagées (mémoires de traduction, bases terminologiques et moteurs de traduction automatique spécialisés préalablement sélectionnés). Il s'agirait d'un outil de TAO sans installation requise et facile à prendre en main. Dans une optique d'amélioration continue, l'objectif serait en effet de mettre en place des processus combinant traduction automatique et TAO avec alimentation de mémoires de traduction et bases terminologiques. Ces ressources pourraient donc être utilisées non seulement pour aider le travail des intervenants dans le processus de traduction, mais aussi pour personnaliser et améliorer les résultats des moteurs de traduction automatique ouverts ou semi-ouverts retenus, même si le niveau et les modalités de personnalisation devront être davantage étudiés ; il peut en effet y avoir des niveaux d'interdisciplinarité plus ou moins importants parmi les différentes disciplines.

Une fois le contenu post-édité et devenu publiable, il serait publié sur le démonstrateur et/ou récupéré par l'éditeur ou la plateforme l'ayant mis à disposition ; le format devrait donc permettre une intégration aisée dans les chaînes éditoriales et leurs flux de travail.

Schéma simplifié

Phase 1

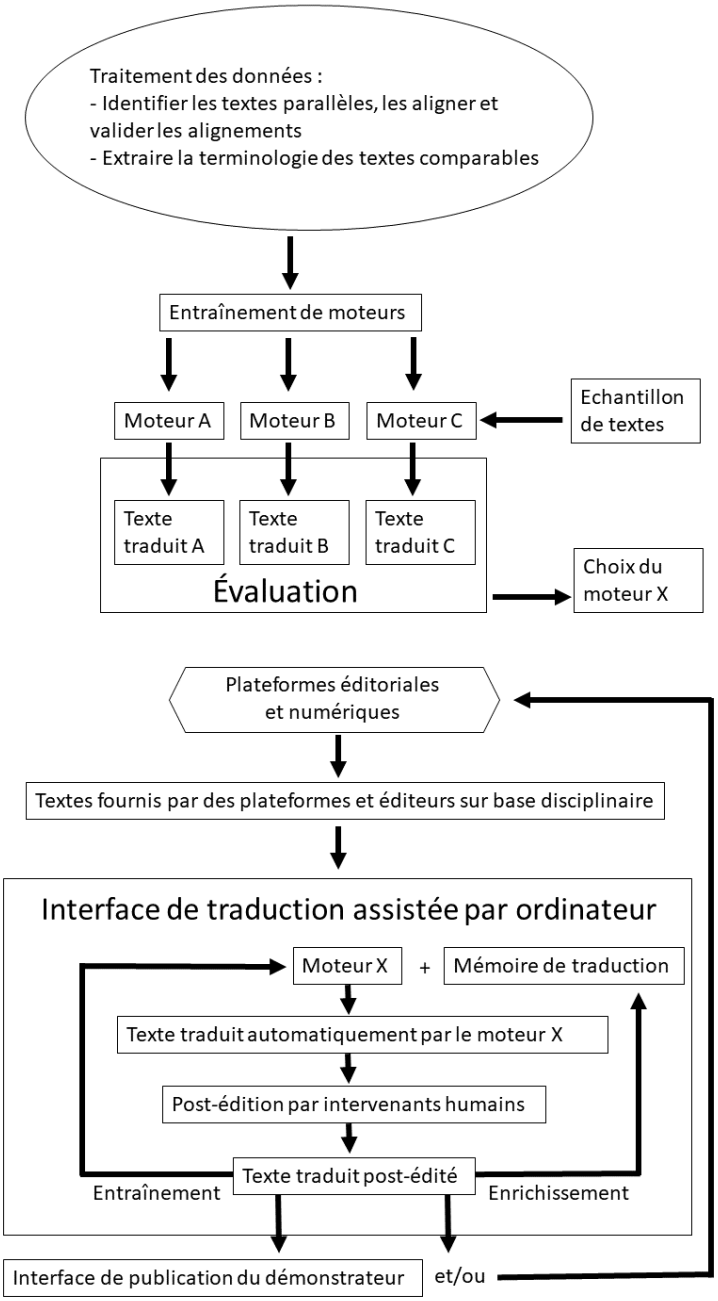
Récupération et traitement de corpus disciplinaires

Phase 2

Évaluation et choix de moteurs de traduction automatique

Phase 3

Démonstrateur : outil mutualisé de traduction de textes scientifiques



Clairement, de nombreuses questions techniques et juridiques se posent, sans oublier les questions liées à l'identification des porteurs de projet et des post-éditeurs, ainsi qu'aux aspects ergonomiques de l'utilisation de la plateforme. Par exemple :

- Quelles autorisations sont requises pour effectuer des traductions ?
- Est-il nécessaire de demander la validation des auteurs avant de publier les traductions ?
- Quels droits d'auteur s'appliquent aux traductions, d'autant plus s'il s'agit d'une post-édition de traduction automatique ?[33]
- Comment prendre en charge automatiquement les formats de publication des différents éditeurs et plateformes ?
- Comment rapatrier les textes traduits vers les sites des éditeurs et des plateformes, le cas échéant ?
- Qui va assurer le travail de post-édition ? Des traducteurs professionnels ? Des chercheurs ? Des binômes constitués de chercheurs et traducteurs ?
- Comment garantir l'interaction entre traducteurs, auteurs, éditeurs et spécialistes du domaine ?
- Quels mécanismes de contrôle qualité mettre en place ?

Ces questions pourront être étudiées dès le début de 2021 en impliquant des experts techniques et juridiques ainsi que les acteurs de la communauté scientifique (plateformes, éditeurs, chercheurs, etc.) afin de prendre en compte leurs attentes et besoins. L'idée n'est pas justement de présenter un produit fini mais plutôt une expérimentation ouverte à tous les acteurs intéressés, qui auront à leur disposition un démonstrateur pour expérimenter et faire des retours sans alourdir leur propre environnement de travail. À cette fin, il sera nécessaire de constituer de nouveaux groupes de travail rassemblant les experts, les éditeurs et les plateformes sur base disciplinaire, sans oublier les chercheurs.

En parallèle, ce projet opérationnel devra être accompagné et complété par des projets de recherche qui permettront de mieux cerner des questions comme, par exemple, la contamination d'une discipline à l'autre, le niveau de spécialisation requis pour les moteurs, etc.

3.2.6 Élaboration d'un guide à destination des chercheurs et des institutions de recherche sur la traduction automatique, la rédaction en langue étrangère et la « rédaction claire » (adaptée à la traduction automatique)

Quels sont les acteurs qui effectueront la post-édition de traduction automatique dans les textes scientifiques et quelles seront les compétences requises ? Ces questions ne sont pas encore résolues et ne donneront sans doute pas lieu à des réponses univoques. Les chercheurs ne seront donc pas les seuls acteurs impliqués dans les travaux de post-édition mis en place dans le cadre du projet Traductions et science ouverte, surtout dans une optique de montée en puissance et de centralisation des processus de traduction. Néanmoins, il convient de commencer d'ores et déjà à les sensibiliser sur les questions liées à la traduction automatique, à la rédaction en langue étrangère et aux principes de « rédaction claire »,

[33] Pascal Reynaud, 2018, Quels droits d'auteur pour le traducteur professionnel ?, Les Éditions de la SFT, Paris, France

essentiels non seulement pour améliorer la compréhension de leurs publications, mais également pour obtenir de bons résultats avec la traduction automatique. Il faudra néanmoins garder à l'esprit les obstacles que ces bonnes pratiques pourront rencontrer : tout d'abord, les revues à fort impact donnent des indications de rédaction assez strictes, qui peuvent parfois aller à l'encontre de la rédaction claire. Ensuite, effectuer une rédaction claire en français, par exemple, risque de ne pas répondre forcément aux exigences de l'article scientifique en français, tout comme sa traduction vers l'anglais pourrait ne pas correspondre aux exigences de l'article scientifique en anglais (modes d'argumentation, formes et temps des verbes, etc.)

Le groupe propose donc de rédiger un guide abordant les points suivants :

- enjeux du multilinguisme et de la science ouverte (meilleur référencement et plus grande visibilité pour les travaux) ;
- inventaire des outils de traduction automatique (types et caractéristiques) ;
- bonnes pratiques pour l'utilisation de la traduction automatique ;
- guide de style pour la rédaction scientifique en anglais/dans d'autres langues choisies, si disponible(s) ou à compiler (afin de donner une référence des attendus en termes de qualité publiable) ;
- principes de « rédaction claire » afin de produire des contenus adaptés, entre autres, à la traduction automatique.

Ces thèmes ont déjà été abordés, par exemple, dans le cadre de certains des projets lauréats de l'appel Traductions scientifiques, le travail de la chercheuse canadienne Lynne Bowker, le livret Rédiger clairement[34] de la Commission européenne, etc. Tous ces travaux pourront servir de base pour la rédaction du guide qui sera un point de départ pour concevoir et prévoir à terme des formations sur mesure, une fois que les outils et les prototypes de processus éditoriaux auront été testés dans le cadre, par exemple, du démonstrateur. Ce travail de sensibilisation permettra sans doute aussi d'obtenir des retours sur les éventuels usages déjà en place. Ce guide pourrait donc faire l'objet de révisions à mesure que l'expérience des communautés impliquées se renforcerait.

3.2.7 Étude de pistes de collaboration dans les réseaux d'éditeurs européens pour la constitution de corpus

Comme déjà évoqué, la constitution de corpus multilingues est un enjeu crucial pour le développement de technologies linguistiques capables de répondre à des besoins spécifiques. Plusieurs institutions, y compris au niveau européen, ont pris conscience que les initiatives exclusivement nationales peuvent rapidement montrer leurs limites et pour cela encouragent la coopération internationale par le biais de réseaux et dispositifs de financement dédiés. Il serait donc opportun de démarrer ou renforcer des collaborations dans l'édition scientifique avec des acteurs en dehors de la France et la Francophonie. Ci-dessous sont listés des appels récurrents sur lesquels se positionner éventuellement à l'avenir :

- Appels à propositions CEF Telecom de la Commission européenne (appel 2020 à titre d'exemple)
- Appels European Language Grid (appel 2020 à titre d'exemple)
- Financement de projets par l'EAMT - European Association for Machine Translation (projets financés à titre d'exemple)

[34] Rédiger clairement: les conseils de Claire
https://ec.europa.eu/info/sites/info/files/clear_writing_tips_fr.pdf [consulté le 27 octobre 2020]

4. Conclusions

Dans le monde de la recherche d'aujourd'hui, de plus en plus internationalisé, la traduction a certainement un rôle essentiel à jouer pour rendre l'actuel système de publication plus équitable et pour élargir l'accès à l'information scientifique à plusieurs niveaux de la société. Certes la traduction a un coût qui n'est pas toujours viable pour les institutions de recherche, mais aujourd'hui il est envisageable d'optimiser les processus de traduction à l'aide de technologies de plus en plus performantes. Il faut néanmoins garder une approche réaliste et raisonnée, tenant compte des spécificités de la communication scientifique, au sens large et dans les différentes disciplines, du degré de maturité des technologies et de ses utilisateurs, ainsi que des contraintes dictées par les ressources disponibles. Afin de construire un modèle qui soit contextuellement pertinent et économiquement durable, il faudra impliquer tous les acteurs intéressés et mener des expériences pilotes innovantes mais rigoureuses, dans le but d'envisager un déploiement à grande échelle fondé sur des ressources et des technologies les plus ouvertes possibles.