De Europa

European and Global Studies Journal

Università degli Studi di Torino

Multilinguisme et variétés linguistiques en Europe à l'aune de l'intelligence artificielle

Multilinguismo e variazioni linguistiche in Europa nell'era dell'intelligenza artificiale

Multilingualism and Language Varieties in Europe in the Age of Artificial Intelligence

Édité par, a cura di, edited by

Rachele Raus, Università di Bologna
Alida Maria Silletti, Università di Bari
Silvia Domenica Zollo, Università di Verona
John Humbley, Université de Paris

UNIVERSITÀ DI TORINO

Special Issue - 2022

De Europa

De Europa
Special Issue 2022

Multilinguisme et variétés linguistiques en Europe à l'aune de l'intelligence artificielle

Multilinguismo e variazioni linguistiche in Europa nell'era dell'intelligenza artificiale

Multilingualism and Language Varieties in Europe in the Age of Artificial Intelligence

Édité par, a cura di, edited by

Rachele Raus, *Università di Bologna*
Alida Maria Silletti, *Università di Bari*
Silvia Domenica Zollo, *Università di Verona*
John Humbley, *Université de Paris*

Special Issue - 2022

De Europa
European and Global Studies Journal
www.deeuropa.unito.it

Collane@unito.it
Università di Torino

ISBN ebook: 9788875902179
ISBN cartaceo: 9788855268431

Ledizioni LediPublishing
Via Antonio Boselli, 10
20136 Milano – Italia
www.ledizioni.it
info@ledizioni.it

With the support of the Erasmus+ Programme of the European Union

Jean Monnet Chair
The EU in a Challenging World

In cooperation whit:

Introduction/ Introduzione / Introduction
Rachele Raus 7

Première partie : réflexions et études de cas

Introduction
Réflexions et études de cas à l'aune de l'intelligence artificielle. Vers de nouveaux observables linguistiques ?
John Humbley, Silvia Domenica Zollo 35

Quelques réflexions sur le multilinguisme à l'aune de l'intelligence artificielle

Intelligence artificielle et langues minoritaires : du bon ménage ? Quelques pistes de réflexion
Giovanni Agresti 47

Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive
Guido Vetere 69

Études de cas

Enabling additional official languages in the EU for 2025 with language-centred Artificial Intelligence
Kepa Sarasola, Itziar Aldabe, Nora Aranberri 91

Langages et savoirs : intelligence artificielle et traduction automatique dans la communication scientifique
Maria Luisa Villa, Maria Teresa Zanola, Klara Dankova 107

Variation et traduction

Terminologie, intelligence artificielle, psychologie cognitive : réflexions sur les interactions possibles dans l'étude de la variation en langue spécialisée
Anne Condamines 131

Human-machine interaction: how to integrate plain language rules in the revision cycles of Neural Machine Translation output
Christopher Gledhill, Maria Zimina 149

A Journey in Neural Machine Translation
Philippe Langlais 173

Deuxième partie : expérimentations pédagogiques

Introduction
Expérimentations pédagogiques : perception et utilisation de l'intelligence artificielle dans la formation universitaire
Alida Maria Silletti, Rachele Raus 199

Le multilinguisme européen et l'IA.
Enquête auprès des futurs décideurs
Dardo de Vecchi 215

Les dispositifs de traduction automatique et la recherche terminologique comme outils pédagogiques pour des étudiant·e·s en droit
Francesca Bisiani 247

Fraseologia, traduzione e *digital literacy* nel contesto universitario: riflessioni e proposte per un percorso didattico sperimentale
Silvia Domenica Zollo, Silvia Calvi 263

Variation terminologique et traduction automatique : une expérience didactique dans l'enseignement du français sur objectif spécifique (FOS)
Jana Altmanova, Luca Bottiglieri 285

Les genres textuels

La traduzione automatica neurale: uno strumento di sensibilizzazione per la formazione universitaria in lingua e traduzione francese
Ilaria Cennamo, Maria Margherita Mattioda 307

Artificial Intelligence and Machine Translation:
perceptions, opinions and experiences of Italian Graduate
Students of English as a Foreign Language
Alessandra Molino 337

Assessing the efficacy of machine translation across genres
Chiara Abbadessa, Monica Albini, Elisa De Paoli, Francesca Del Nobile 355

Intelligenza artificiale e traduzione automatica nel contesto della
formazione universitaria di lingua tedesca
Lucia Cinato 365

Annexes 385

De Europa Special Issue 2022

Introduction

Rachele Raus

Ce numéro spécial de la revue *De Europa* a été élaboré à partir d'une réflexion interdisciplinaire et multilingue qui a été menée dans le cadre d'une recherche sur les droits, le multilinguisme et les variétés linguistiques en Europe à l'aune de l'IA[1] que je coordonne à l'intérieur du projet *Artificial Intelligence for European Integration*[2], dirigé par Umberto Morelli, promu par le Centre d'études européennes *To-EU* de l'Université de Turin[3] et co-financé par la Commission de l'Union européenne.

Notre propos était de réfléchir plus généralement sur les conséquences négatives et/ou positives de l'IA sur les variétés linguistiques et le multilinguisme, ce dernier étant une valeur de l'UE. La recherche a débuté le 6 octobre 2020 lors de l'intervention de Laurent Romary[4], directeur du Comité 37 de l'Organisation internationale de normalisation (ISO), à l'intérieur d'un premier colloque international organisé en distanciel à l'Université de Turin par le groupe pilote du projet. Ensuite, le 23 avril 2021, nous avons eu l'occasion d'organiser un deuxième colloque en distanciel portant sur la thématique spécifique de notre intérêt. Les interventions présentées à ce colloque sont désormais disponibles dans le site du projet[5]. La journée a été suivie par une matinée de séminaire (*workshop*) où les participant.e.s venant de plusieurs universités italiennes et françaises ont eu l'occasion de présenter leurs travaux inédits sur la didactique des langues étrangères et l'IA.

Rachele Raus, Università di Bologna, rachele.raus@unibo.it

[1] Cette recherche a regroupé jusqu'à aujourd'hui environ une centaine de personnes venant de plusieurs pays de l'UE, du Canada et du Brésil.

[2] Voir le lien http://www.jmcoe.unito.it/home (dernière date de consultation de tous les sites cités : le 20 octobre 2021).

[3] https://www.dcps.unito.it/do/home.pl/View?doc=toeu.html

[4] https://media.unito.it/?content=9806

[5] http://www.jmcoe.unito.it/

Compte tenu de l'intérêt du débat soulevé, nous avons décidé de recueillir quelques-unes des contributions des deux journées du 23 et 24 avril 2021, dûment retravaillées à la suite des nombreuses rencontres des groupes de travail et des personnes concernées, et d'y ajouter des essais de collègues qui ont voulu participer à cette initiative, en enrichissant un débat déjà passionnant. Par conséquent, ce que nous proposons par ce numéro est un livre inédit qui peut se vanter d'être parmi les premiers à s'occuper de ce type de thématique, du moins en Europe.

1. La question initiale

Dès les années 1990, plusieurs organisations internationales, notamment l'Unesco[6], ont commencé à se poser la question de la « fracture numérique » causée, entre autres, par les technologies actuelles qui sont plus diffusées dans certains pays que dans d'autres. Sans compter d'autres risques du même genre : par exemple, le fait que les langues orales ont moins la possibilité de devenir visibles dans la toile ; que les grands ensembles de corpus et des données (*big data*) sont de plus en plus gérés par des entreprises privées[7] ; que certains pays, comme le Canada ou, en Europe, la France et l'Allemagne, ont investi massivement dans la production de pages publiées dans la toile et également dans l'industrie des langues ; que la plupart des données linguistiques et des corpus sont normalement disponibles en langue anglaise ; que souvent les outils informatiques multilingues utilisent l'anglais comme langue pivot (voir la contribution de Vetere dans ce numéro), etc.

Les groupes de travail[8] que nous avons organisés autour de ces questionnements ont ouvert plusieurs pistes de recherche possibles :

1. la question des droits linguistiques par rapport à la possibilité de certains pays d'investir dans l'industrie des langues aux dépens des autres. Cette thématique, soulevée par des linguistes et des juristes des groupes de travail, a été abordée par le Rapporteur spécial des Nations unies sur les questions relatives aux minorités, Fernand de Varennes[9],

[6] Cf. http://www.unesco.org/new/index.php?id=50219

[7] A propos de la gestion des données par rapport à l'IA, cf. l'intervention de l'informaticien Ciro Cattuto au colloque du 6 octobre 2020 au lien : https://media.unito.it/?content=9806

[8] Les groupes ont été divisés par langue véhiculaire (FR, IT, EN).

[9] https://drive.google.com/file/d/1YPSFX-v9ITB4N1tj_eFOl7Sinc9NBgcY/view

lors du colloque du 24 avril 2021. Le débat se poursuivra en 2022, lors d'un colloque organisé par l'Université Catholique de Lille et prévu pour le printemps de l'an prochain;

2. la nécessité de sélectionner attentivement les sources et les données utiles pour entraîner les algorithmes d'IA.

Cette deuxième question en soulève deux autres qui lui sont étroitement liées : d'abord, la relation que ces données entretiennent avec les formes de discrimination, ce que Bartoletti a appelé la « discrimination algorithmique » (Bartoletti 2020) ; ensuite, la relation entre ces données et des sources « d'autorité », ce dont on a commencé à discuter avec Laurent Romary lors du colloque du 6 octobre 2020. En effet, le critère de source d'autorité proposé par l'ISO (2009 : 7) recommande que cette source soit officielle ou garantisse des critères de qualité, ce qui implique, de manière indirecte, la présence d'une source institutionnelle. C'est sans doute l'une des raisons pour laquelle les tous premiers entraînements des algorithmes neuronaux utilisés dans les logiciels de traduction automatique ou dans les dispositifs numériques ont été faits sur les corpus multilingues des organisations internationales. Une autre en est que celles-ci disposent d'une grande quantité de données multilingues déjà disponibles et prétraitées pour le traitement automatique. C'est ce qui est arrivé pour le traducteur de Google[10], dont les réseaux neuronaux d'apprentissage profond se sont entraînés sur les corpus des Nations unies (ONU 2017 : 120), ainsi que pour Facebook, qui a utilisé les documents multilingues du Parlement européen pour entraîner ses traducteurs automatiques (Le Cun 2019 : 277), ou encore pour le concordancier Linguee[11], dont les corpus fournissent le matériel d'entraînement au traducteur automatique DeepL[12], corpus qui viennent des institutions de l'UE, de l'ONU, du gouvernement canadien… Cela dit, l'utilisation de corpus et de données plus variés, mais non institutionnels, qui donc ne sont pas forcément vérifiés ni forcément de qualité, a fini par améliorer les performances des outils informatiques par rapport à d'autres qui utilisent des corpus institutionnels.

Pendant le colloque d'octobre, nous avons justement cité l'exemple de l'accord des noms et des adjectifs au féminin. En effet, ce problème, qui

[10] https://translate.google.com
[11] https://www.linguee.fr
[12] https://www.deepl.com/fr/translator

caractérise avant tout les noms de métiers et de profession, s'aggrave lors de la rédaction ou de la traduction automatique des textes faite par des outils dont les réseaux de neurones d'apprentissage profond s'entraînent sur des corpus institutionnels qui, privilégiant des textes juridiques et/ou politiques, finissent par utiliser le « neutre » masculin pour renvoyer aux catégories d'acteurs. Cela produit des erreurs d'accord du féminin lors de la traduction automatique des langues romanes, comme il arrive pour le traducteur de Google ou DeepL, qui sont parmi les outils les plus utilisés, entre autres, par les universitaires italiens (voir les contributions dans la deuxième partie de ce numéro).

Ce problème pourrait être résolu, du moins en partie[13], en utilisant des corpus variés qui ne sont pas forcément vérifiés, et donc non institutionnels, qui seraient plus proches des usages linguistiques et de la langue courante.

Le questionnement sur les données et sur leurs critères de sélection a conduit à s'interroger sur la possibilité de prévoir des critères en amont de l'entraînement des réseaux neuronaux. L'intervention humaine par rapport à la « machine » se déplacerait donc du travail actuel (en aval) de révision des textes produits par celle-ci au travail préalable (en amont) de réflexion et d'implémentation des critères linguistiques et discursifs utiles pour bien entraîner l'IA (sélection de sous-corpus choisis exprès pour la tâche spécifique ; mise en place de contraintes qui permettraient d'améliorer et de faciliter l'apprentissage supervisé de la machine, etc.). Ces critères permettraient de réaliser concrètement les recommandations du livre blanc européen de 2020 par rapport au contrôle humain lors de la conception de l'outil afin d'éliminer des discriminations linguistiques et de genre et d'« imposer des contraintes opérationnelles » favorisant le multilinguisme et les variations linguistiques (Commission européenne 2020 : 22 et 25).

Il s'agit sans aucun doute d'une solution onéreuse, qui demande une réflexion préalable sur ce qui fait « autorité » par rapport à l'entraînement de l'IA et qui implique également tout un travail d'étiquetage de corpus de

[13] Bien que le manque de féminisation de certains noms de métiers et de profession caractérise plus généralement plusieurs langues romanes, entre autres le français et l'italien, cette tendance empire dans les corpus institutionnels qui entraînent les algorithmes et qui recourent normalement au masculin « neutre » pour renvoyer aux acteurs dans les discours juridiques et ensuite administratifs (voir aussi Marzi 2021).

grande taille pour les rendre plus utiles par rapport à l'apprentissage. Cela dit, il s'agit d'un travail lourd mais nécessaire si l'on veut que l'IA devienne réellement un outil au service du multilinguisme. En effet, ce n'est pas parce qu'un traducteur automatique est multilingue qu'il peut favoriser le multilinguisme si, par exemple, il utilise l'anglais comme langue pivot, comme il arrive pour la plupart des outils de l'industrie des langues actuelle.

Outre la question des droits et des données, les groupes de travail sur le multilinguisme et l'IA se sont aussi intéressés à un volet pédagogique de la recherche, ce qui a permis de promouvoir des travaux en classe centrés sur les langues et l'IA à travers l'utilisation d'outils de traduction automatique exploitant des réseaux neuronaux d'apprentissage profond (voir la deuxième partie de ce numéro).

2. D'une réflexion critique à une IA qui favorise le multilinguisme

Bien que plusieurs contributions de ce numéro finissent par être critiques par rapport aux conséquences de l'IA sur le multilinguisme (voir, entre autres, les articles d'Agresti et de de Vecchi dans ce numéro) et par rapport à un marché, tel celui de l'industrie des langues, qui aujourd'hui donne une fausse perception des langues comme de simples moyens de communication, voire des « codes » interchangeables, il faut admettre que c'est justement en questionnant l'IA que nous avons pensé à des possibilités concrètes de l'utiliser de manière différente et donc en faveur du multilinguisme.

Cela nous a permis de revenir sur la question des données de manière nouvelle, en proposant de concevoir et de fabriquer une application multilingue qui faciliterait la communication inclusive et qui serait destinée aux administrations publiques. Cette application serait d'abord conçue pour la langue italienne et ensuite pour d'autres langues européennes (en français, puis en allemand et enfin en espagnol et/ou en anglais, en tenant compte de la variation diatopique de ces langues). Elle sera réalisée avec le soutien de l'École polytechnique de Turin à partir de critères qui puissent développer une IA en faveur des variétés linguistiques et du multilinguisme. Nous espérons présenter l'application italienne lors d'une journée d'études que nous souhaitons organiser pour la fin de 2023 et qui permettrait de présenter également des initiatives similaires à la nôtre,

comme le nouveau traducteur automatique que Facebook a inauguré en 2020 et qui n'utilise plus l'anglais comme langue pivot[14], ou comme le transcripteur IA Azur de Microsoft qui est utilisé par le Parlement européen et qui tient compte des différents accents régionaux ou nationaux (Aeles 2020). Cette journée montrerait la possibilité d'améliorer les performances de l'IA grâce au choix de critères discursifs et linguistiques qui tiennent compte des variations diaphasiques, diatopiques et diastratiques des langues.

Par rapport au volet pédagogique, les groupes de travail ont mis en place, pendant l'année académique 2020-2021, une expérimentation dont les résultats sont présentés dans la deuxième partie de ce numéro. Cette expérimentation a permis de mettre en relief comment une formation ciblée permet de développer une conscience critique par rapport à l'IA, ce qui favorise une meilleure utilisation des outils de l'industrie des langues qui s'appuient sur l'IA.

Cette réflexion ne porte pas atteinte à la confiance initiale que les étudiant.e.s déclarent faire à l'IA, mais tend à la renforcer grâce à une meilleure compréhension de ces outils et de leur potentiel.

Au final, ce numéro de la revue *De Europa* entend favoriser l'amélioration de l'IA d'une part, en privilégiant des critères linguistiques et discursifs précis fixés en amont de son apprentissage et d'autre part, en permettant une compréhension majeure de la manière dont elle privilégie certaines données de sortie. De cette manière, les articles ci-rassemblés entendent également contribuer aux produits de recherche finaux du projet, qui sont attendus pour 2023, à savoir :

- des recommandations à l'intention des décideurs européens pour l'élaboration de politiques qui favorisent la mise en œuvre d'une intelligence artificielle respectueuse du multilinguisme ;
- un vade-mecum, destiné au personnel travaillant dans des domaines professionnels (technique, informatique, organismes de normalisation, ...), qui contiendra des suggestions pour concevoir des algorithmes d'IA qui favorisent le multilinguisme et les variations diatopiques et diastratiques des langues, en se focalisant notamment sur les langues officielles de l'UE.

[14] Cf. https://www.20minutes.fr/high-tech/2889599-20201020-facebook-met-point-outil-traduction-automatique-entre-100-langues-differentes

3. Présentation des contributions

Ce numéro, qui contient des contributions trilingues (IT, FR, EN) résumant les points forts de la première année de recherche des groupes de travail, se pose comme le premier d'une trilogie, chaque livre renvoyant aux résultats d'une année de travail. Nous avons voulu privilégier une approche de vulgarisation des contenus scientifiques de manière à joindre un public très large, sans rien enlever à la rigueur scientifique de nos propos.

De même, nous avons essayé de favoriser une écriture inclusive, sans pour autant obliger les personnes qui ont participé à cette publication à privilégier des choix linguistiques spécifiques. C'est la raison pour laquelle, dans certaines contributions, on a privilégié la féminisation (par le point médian, par les tirets ou par des parenthèses) et que dans d'autres (la majorité des cas) on a choisi d'utiliser le masculin soi-disant « neutre ».

Par rapport à la structure du numéro, nous l'avons divisé en deux parties : dans la première, nous avons voulu présenter des réflexions générales sur l'IA par rapport au multilinguisme et aux variétés linguistiques. Ensuite, nous avons regroupé les contributions qui portent sur des études de cas spécifiques, surtout des cas concernant la traduction automatique et l'IA, qui constituent la toute dernière sous-section de la première partie.

Dans la seconde section du numéro nous avons rassemblé les articles concernant les résultats des expérimentations pédagogiques menées par des universités et des écoles professionnelles françaises et italiennes qui ont participé au projet : la Kedge Business School, l'EDEH, l'EM-Normandie, la Montpellier Business School et l'École supérieure de commerce de Clermont, l'Université Catholique de Lille, l'Université Catholique du Sacré-Cœur de Milan, l'Université de Naples — L'Orientale, l'Université de Turin et l'Université de Vérone. Nous avons également pu intégrer à cette section des articles ou des propos venant des étudiant.e.s universitaires (niveau Master 1-2 ou 3^{e} cycle), étant donné que le projet prévoyait une méthode de recherche-action.

Avant de laisser nos lectrices et nos lecteurs à la lecture du numéro, nous tenons à remercier John Humbley, Alida Maria Silletti et Silvia Domenica Zollo pour leur travail infatigable et ponctuel de relecture et de mise en forme des articles, ainsi que pour avoir nourri la réflexion commune qui a permis à ce travail de voir le jour.

Nous remercions également Umberto Morelli, qui coordonne le projet *Artificial Intelligence for European Integration* et qui a toujours su gérer au mieux tous les groupes de travail, même lorsque les conditions pour mener à bien le projet sont devenues difficiles à cause de la pandémie à COVID-19.

Enfin, merci à toutes les personnes qui ont contribué de manière plus ou moins directe à nourrir le débat sur ces thématiques passionnantes.

Nous espérons qu'il pourra se poursuivre et devenir de plus en plus riche, l'IA étant désormais devenue un élément incontournable de notre vie quotidienne.

Bibliographie

Aeles Joris (2020). *Rapprocher le Parlement européen des citoyens grâce à l'IA*.

https://pulse.microsoft.com/fr-be/transform-fr-be/government-fr-be/fa1-rapprocher-le-parlement-europeen-des-citoyens-grace-a-lia/

Bartoletti Ivana (2020). *An Artificial Revolution. On Power, Politics and AI*. Edimbourg : Indigo.

Commission européenne (2020). *Livre blanc. Intelligence artificielle. Une approche européenne axée sur l'excellence et la confiance*. COM(2020) 65 final.

Le Cun Yann (2019). *Quand la machine apprend*. Paris : Odile Jacob.

Marzi Eleonora (2021). « La traduction automatique neuronale et le biais de genre : le cas des noms de métiers entre l'italien et le français ». *Synergies Italie*, 17, 19-36. https://gerflint.fr/Base/Italie17/marzi.pdf

Organisation des Nations unies (2017). *Change. Rapport annuel 2017*. Genève : Nations unies.

Organisation internationale de normalisation — ISO (2009). *Norme 23185. Assessment and benchmarking of terminological resources*. Genève : ISO.

De Europa Special Issue 2022

Introduzione

Rachele Raus

Il presente numero speciale di *De Europa* è il frutto di una riflessione interdisciplinare e multilingue maturata attorno a diversi eventi organizzati nell'ambito del panel concernente i diritti e le variazioni linguistiche in Europa nell'era dell'intelligenza artificiale[1] da me diretto all'interno del progetto *Artificial Intelligence for European Integration*[2], promosso dall'Università di Torino (direzione del prof. Umberto Morelli) e cofinanziato dalla Commissione europea.

L'interrogativo iniziale che abbiamo voluto sollevare è se l'IA potesse avere un impatto negativo sulle varietà linguistiche e sul multilinguismo, valore "aggiunto" dell'UE[3], o se potesse, e in che modo, divenire utile per la promozione di essi. La ricerca ha dato modo di organizzare un primo convegno a distanza il 6 ottobre del 2020 al quale è stato invitato il prof. Laurent Romary[4] quale rappresentante del Comitato tecnico 37 dell'ISO e successivamente un secondo convegno, svoltosi a distanza il 23 aprile 2021, i cui interventi sono disponibili nel sito del progetto[5] e al quale è seguita una giornata seminariale (*workshop*) in cui sono stati presentati dei lavori inediti sulla didattica e l'IA svolti in numerose università italiane e francesi.

Dato l'interesse del dibattito sollevato, si è scelto di raccogliere alcuni dei contributi proposti in queste giornate, opportunamente rielaborati

Rachele Raus, Università di Bologna, rachele.raus@unibo.it

[1] Il panel ad oggi ha coinvolto un centinaio di persone circa provenienti da vari paesi dell'UE, dal Canada e dal Brasile. Questo senza considerare il contributo fornito dalla componente studentesca che è stata coinvolta nelle ricerche sulla didattica.

[2] Maggiori dettagli al link http://www.jmcoe.unito.it/home (tutti i link della presentazione sono stati consultati il 20 ottobre 2021).

[3] Così è infatti spesso definito proprio nei siti istituzionali europei, in particolare del Parlamento.

[4] https://media.unito.it/?content=9806

[5] http://www.jmcoe.unito.it/

anche a seguito dei numerosi incontri avuti con molte delle persone che vi hanno partecipato, integrandoli con articoli di colleghe e colleghi che hanno voluto arricchire ulteriormente un dibattito già fruttuoso. Conseguentemente, ne è risultato un volume interamente inedito, che può dirsi tra i primi ad affrontare, almeno in Europa, questo tipo di tematiche.

1. L'interrogativo iniziale

Dagli anni 1990, diverse organizzazioni internazionali, e in particolare l'Unesco[6], hanno cominciato ad affrontare il problema del divario di conoscenza tra i vari paesi a seguito del fatto che le tecnologie attuali, a cominciare da Internet, siano più diffuse in alcuni paesi piuttosto che in altri. La presenza di lingue orali che hanno meno la possibilità di divenire visibili nel web, la gestione dei grandi dati, anche linguistici, che sono sempre più in mano a colossi privati[7], la capacità e la volontà di alcuni paesi come il Canada, o, in Europa, la Francia o la Germania di investire massivamente nella produzione di pagine web e nei prodotti dell'industria linguistica rispetto ad altri che hanno meno disponibilità economiche, la disponibilità maggiore dei dati in lingua inglese, la concettualizzazione della traduzione multilingue a partire dall'inglese veicolare (cfr. il contributo di Vetere nel presente volume), ecc. sono tutte questioni che certamente si ripercuotono sul multilinguismo e sulla questione delle lingue minoritarie che sarà trattata in questo volume.

Dalle prime riflessioni dei gruppi di lavoro[8] che abbiamo costituito attorno a queste tematiche sono scaturiti cinque assi di ricerca privilegiati che abbiamo voluto ulteriormente indagare. Il primo concerne la capacità dei paesi di investire nell'industria linguistica e più generalmente le politiche linguistiche implementate dai diversi paesi dell'UE per rispettare i diritti linguistici. Questo primo asse, che ha coinvolto persone specialiste nell'ambito linguistico e giuridico, è stato introdotto dall'intervento del prof. Fernand de Varennes[9], relatore speciale delle Nazioni Unite, nell'ambito dei lavori del 24 aprile 2021 e proseguirà con i lavori promossi dall'U-

[6] Cfr., ad esempio, http://www.unesco.org/new/index.php?id=50219

[7] A proposito della gestione dei *big data* in relazione all'IA, cfr. l'intervento di Ciro Cattuto al convegno del 6 ottobre 2020, disponibile al link https://media.unito.it/?content=9806

[8] I gruppi di lavoro sono stati suddivisi sulla base delle lingue veicolari utilizzate (FR, IT, EN).

[9] https://drive.google.com/file/d/1YPSFX-v9ITB4N1tj_eFOl7Sinc9NBgcY/view

niversité Catholique di Lille che sta attualmente organizzando un convegno su tali tematiche per la primavera del 2022.

Il secondo asse di ricerca mira a riflettere sulla necessità di selezionare con attenzione le fonti e i dati utili per l'addestramento dell'IA, e si correla direttamente ad altri due degli assi proposti, quello sulla relazione tra tali dati e la "discriminazione algoritmica" (Bartoletti 2020) e quello sulla relazione tra i dati e la nozione di fonte "autorevole", sulla quale si è cominciato a discutere con il prof. Romary durante il convegno del 6 ottobre 2020. In effetti, il criterio della fonte autorevole dell'ISO (ISO 2009: 7) raccomanda che tale fonte sia ufficiale o garantisca criteri di qualità, cosa che implica indirettamente la presenza di una fonte di tipo istituzionale. Questo è probabilmente uno dei motivi per cui, assieme al fatto di disporre di grandi dati utili, i primi addestramenti degli algoritmi neurali di apprendimento profondo utilizzati per i *tool* di traduzione automatica o nell'industria linguistica si sono basati sui corpora delle organizzazioni internazionali. Così è avvenuto per il traduttore di Google[10], rispetto alle fonti delle Nazioni Unite (ONU 2017: 120), per gli algoritmi di traduzione automatica utilizzati da Facebook rispetto alla documentazione del Parlamento europeo (Le Cun 2019: 277), per l'addestramento di software di concordanze come Linguee[11], i cui corpora forniscono il materiale per il traduttore automatico DeepL[12], rispetto ai testi del governo canadese e/o delle istituzioni dell'UE e dell'ONU… Tuttavia, l'utilizzo di corpora e dati di tipo misto, non per forza istituzionale né per forza verificati e perciò di qualità, ha finito per dare risultati migliori rispetto alle performance dell'IA utilizzata in questi strumenti. Proprio durante il convegno di ottobre, citavamo, ad esempio, il fatto che la variante di genere (es. la declinazione al femminile), già problematico più generalmente nell'italiano e nelle lingue romanze soprattutto in relazione ai titoli e nomi di professione, viene ulteriormente a perdersi in testi frutto di elaborazione (redazione o traduzione) di strumenti supportati dall'IA addestrata su corpora istituzionali, essenzialmente a carattere politologico-giuridico, che utilizzano massivamente il maschile "neutro" o "non marcato" per indicare delle categorie di attori. Non c'è da sorprendersi, perciò, che testi declinati al femminile in francese diventino esclusivamente maschili nella traduzione ita-

[10] https://translate.google.it/?hl=it
[11] https://www.linguee.it/
[12] https://www.deepl.com/it/translator

liana come ancora avviene per traduttori automatici diffusi come Google o DeepL, peraltro tra i più usati nel mondo universitario italiano come si può evincere dalla seconda parte del presente volume. Questo potrebbe essere evitato, almeno in parte[13], in caso di utilizzo di corpora misti non per forza verificati, e non istituzionali, ma più indicativi rispetto agli usi linguistici della lingua comune.

Questi interrogativi sui dati e sui loro criteri di selezione hanno portato a promuovere la possibilità di inserire correttivi e criteri a monte dell'apprendimento delle reti neurali, non limitando perciò l'intervento umano alle attività di *postediting* dei testi prodotti in modo automatico, come avviene per lo più oggigiorno, ma prevedendolo a monte rispetto all'apprendimento della macchina. In tal senso, i correttivi consentirebbero di realizzare concretamente i criteri indicati nel libro bianco europeo del 2020 di una sorveglianza durante la fase di progettazione dell'IA per eliminare a monte possibilità di discriminazione linguistica e di genere (Commissione europea 2020: 21) e d'implementare "regole operative" (*ibidem*: 24) atte a promuovere il multilinguismo e la variazione linguistica.

Certamente, si tratta di una via più lunga da percorrere, che porta a riflessioni sui corpora "autorevoli" da selezionare e soprattutto sui criteri da adottare in fase di apprendimento supervisionato, nonché implica la necessità di annotare corpora ampi perché diventino abbastanza significativi da poter essere utili per l'addestramento dell'IA, ma si tratta anche di un lavoro necessario se si vuole davvero che essa diventi un vettore a sostegno del multilinguismo. Precisiamo, infatti, che un traduttore automatico multilingue non è di per sé, solo per il fatto di essere multilingue, un supporto al multilinguismo se processa i dati favorendo, ad esempio, l'inglese veicolare che resta maggioritariamente lingua ponte nell'industria linguistica attuale.

Infine, il quinto asse di ricerca si è focalizzato più specificatamente sulla didattica e ha permesso di promuovere in classe lavori centrati sulle lingue e l'IA tramite l'ausilio di strumenti di traduzione automatica basati su algoritmi di apprendimento profondo e quindi con intelligenza artificiale.

[13] Sebbene la mancanza di femminilizzazione di molti titoli e nomi di professione caratterizzi più generalmente le lingue romanze, tra le quali il francese e l'italiano, tale tendenza peggiora nei corpora istituzionali che addestrano gli algoritmi e che ricorrono solitamente al maschile "non marcato" o "neutro" per rinviare agli attori nel discorso giuridico e amministrativo (cfr. anche Marzi 2021).

2. Da una riflessione critica all'IA come vettore a sostegno del multilinguismo

Malgrado alcuni degli interventi presenti in questo volume siano animati da uno spirito di analisi critica rispetto agli attuali utilizzi degli algoritmi nell'industria linguistica e a un mercato delle lingue che spinge inesorabilmente verso una concettualizzazione di esse come di meri strumenti di comunicazione (cfr. gli articoli di Agresti e di de Vecchi all'interno del numero) e quindi veri e propri codici privi di spessore culturale e simbolico, va detto che proprio da tale analisi è sorta la possibilità concreta di un utilizzo diverso e virtuoso dell'IA a favore di un multilinguismo reale.

In tal senso, ad esempio, le ricerche effettuate dai nostri gruppi sugli assi concernenti i dati e il loro utilizzo per l'addestramento di reti neurali stanno confluendo verso la concettualizzazione e la realizzazione di un applicativo utile alla comunicazione inclusiva e rivolto alle pubbliche amministrazioni. Tale applicativo, che sarà dapprima realizzato per la lingua italiana, poi anche per altre lingue europee (anzitutto il francese, a seguire il tedesco e probabilmente anche l'inglese e/o lo spagnolo), sarà sviluppato con il contributo del Politecnico di Torino[14], a partire da criteri linguistici e discorsivi che possano favorire lo sviluppo di un'IA a sostegno della variazione linguistica e del multilinguismo. Il nostro desiderio è che questa iniziativa possa esser presentata in una giornata di studi assieme ad altre consimili, come ad esempio quella del nuovo traduttore automatico inaugurato da Facebook[15] nel 2020 e che non utilizza più l'inglese ponte o come il trascrittore IA Azur di Microsoft che è usato dal Parlamento europeo e che tiene conto degli accenti regionali o nazionali (Aeles 2020). Tali iniziative mostrerebbero come, partendo da criteri discorsivi e attenti alla variazione delle lingue, *in primis* quelle diatopica e diastratica, sia possibile migliorare le performance dell'IA e favorirne conseguentemente un utilizzo realmente virtuoso a sostegno del multilinguismo e della variazione linguistica.

Parimenti, rispetto all'asse cinque, l'esperimento didattico intrapreso nel corso dall'A.A. 2020-2021 ha permesso di mettere in luce come tramite la didattica sia possibile far maturare un senso di riflessione critica nei confronti dell'IA che porti a un utilizzo migliore degli strumenti dell'industria linguistica supportati da reti neurali. Tale riflessione non mina quel

[14] Il team di sviluppo è coordinato dalla prof.ssa Tania Cerquitelli, che ringraziamo.

[15] Cfr. https://www.20minutes.fr/high-tech/2889599-20201020-facebook-met-point-outil-traduction-automatique-entre-100-langues-differentes

senso di fiducia iniziale nei confronti dell'IA che la componente studentesca spesso nutre, come vedremo nella seconda parte del presente volume, ma va a rafforzarlo a fronte di una comprensione maggiore di essa e delle sue potenzialità.

Il presente numero di *De Europa*, quindi, vuole contribuire alla riflessione sull'IA al fine di favorire le sue capacità a supporto del multilinguismo, e ciò grazie da un lato all'addestramento a partire da criteri linguistici e discorsivi precisi stabiliti a monte, dall'altro, consentendo a un pubblico giovane e/o non esperto di acquisire una maggior comprensione e consapevolezza del perché essa finisca per privilegiare determinati output. In tal senso, il volume vuol essere un tassello verso l'elaborazione dei due prodotti finali del progetto che sono attesi per il 2023, ovvero una guida a destinazione dei decisori e delle deciditrici dell'UE rispetto alle politiche multilingui da adottare in relazione all'IA e il vademecum a destinazione di un pubblico professionale (enti di normazione, attori dell'industria linguistica…) per sviluppare strumenti e quadri normativi capaci di favorire un'IA a reale supporto della variazione linguistica e del multilinguismo.

3. La strutturazione del numero

Il presente numero speciale presenta contributi trilingui (IT, FR, EN) relativi al primo anno di ricerca dei gruppi di lavoro. Si tratta del primo volume di una trilogia che nei prossimi anni riassumerà i risultati principali delle ricerche svolte nel triennio.

Quanto agli articoli qui raccolti, si è volutamente privilegiato un approccio di tipo divulgativo in modo che essi, seppur supportati da argomentazione e prove scientifiche, potessero rivolgersi a un pubblico vasto. Per lo stesso motivo, si è cercato di favorire una scrittura che fosse inclusiva, senza però obbligare coloro che hanno contribuito al numero a operare scelte precise e univoche in tal senso. Per questo, in alcuni contributi si è scelto di femminilizzare usando il punto mediano, le parentesi o i trattini, in altri (la maggior parte degli articoli) si è lasciato il maschile "non marcato".

Abbiamo deciso di strutturare il testo in due sezioni: nella prima, abbiamo introdotto una riflessione generale sull'IA rispetto alla variazione linguistica e al multilinguismo (cfr. i contributi di Agresti e di Vetere) e abbiamo illustrato dei casi di studio specifici (cfr. i contributi di Sarasola

et alii e di Villa *et alii*). Entrando sempre più all'interno di casi di studio, l'analisi della traduzione automatica è stata quella che ha maggiormente destato l'attenzione degli addetti e delle addette ai lavori (cfr. gli articoli di Condamines, Gledhill e Zimina, Langlais).

Nella seconda parte del volume, abbiamo raccolto gli articoli concernenti la sperimentazione didattica condotta dalle università e scuole professionali che hanno partecipato alla nostra iniziativa (per questo volume: la KEDGE Business School, l'EDEH, l'EM-Normandie, la Montpellier Business School et l'École supérieure de commerce de Clermont, l'Université Catholique de Lille, l'Università Cattolica del Sacro Cuore di Milano, le Università L'Orientale di Napoli, le Università di Torino e di Verona). In tale sezione, compaiono anche direttamente o indirettamente i contributi di studentesse e studenti che hanno partecipato al progetto (livello magistrale e dottorato), dato che esso prevedeva un processo virtuoso di ricerca-azione e di formazione-azione che producesse ritorni anche dal basso.

Prima di lasciare alla lettura degli articoli, teniamo a ringraziare il collega John Humbley e le colleghe Alida Maria Silletti e Silvia Domenica Zollo dell'eccellente lavoro di curatela del presente volume, nonché per la riflessione comune che ha consentito di realizzarlo nelle migliori condizioni possibili. Un ringraziamento va anche al prof. Umberto Morelli, coordinatore del progetto *Artificial Intelligence for European Integration* che ha dato sempre supporto a ogni iniziativa e ha saputo gestire al meglio ogni difficoltà, non ultima, quella della pandemia da COVID-19 che ha cominciato a infuriare proprio nel periodo in cui il progetto ha preso avvio.

Un ringraziamento sentito va anche a tutte le persone che in modo più o meno diretto hanno contribuito alle ricerche. Senza di loro questo primo volume non sarebbe stato possibile.

Soprattutto, ci auguriamo di poter continuare a nutrire grazie a loro la riflessione comune su una tematica, come quella dell'IA, che è diventata e diventerà sempre più parte del nostro quotidiano.

Bibliografia

Aeles Joris (2020). *Rapprocher le Parlement européen des citoyens grâce à l'IA*.

https://pulse.microsoft.com/fr-be/transform-fr-be/government-fr-be/fa1-rapprocher-le-parlement-europeen-des-citoyens-grace-a-lia/

Bartoletti Ivana (2020). *An Artificial Revolution. On Power, Politics and AI*. Edimburgo: Indigo.

Commissione europea (2020). *Libro bianco sull'intelligenza artificiale — Un approccio europeo all'eccellenza e alla fiducia*. COM(2020)65 final.

International Organization for Standardization — ISO (2009). *Norma 23185. Assessment and benchmarking of terminological resources*. Ginevra: ISO.

Le Cun Yann (2019). *Quand la machine apprend*. Parigi: Odile Jacob.

Marzi Eleonora (2021). "La traduction automatique neuronale et le biais de genre : le cas des noms de métiers entre l'italien et le français". *Synergies Italie*, 17, 19-36. https://gerflint.fr/Base/Italie17/marzi.pdf

Organizzazione delle Nazioni Unite (2017). *Change. Rapport annuel 2017*. Ginevra: Nazioni Unite.

De Europa Special Issue 2022

Introduction

Rachele Raus

This special issue of *De Europa* is the outcome of an interdisciplinary multilingual reflection carried out on research into linguistic rights, multilingualism and language varieties in Europe in the age of artificial intelligence[1]. It is part of the *Artificial Intelligence for European Integration* project[2] I have been coordinating, headed by Umberto Morelli, with the backing of the Centre of European Studies To-EU of the University of Turin[3] and co-financed by the European Commission.

Our aim was to investigate more generally the negative and/or positive outcomes of AI on language varieties and multilingualism, the latter a key value for the EU. The research started on 6 October 2020 with the paper given by Laurent Romary[4], in charge of the Technical committee 37 of the International Standards Organisation (ISO) at the first international conference organised remotely by the University of Turin by the pilot group. Then, on 23 April 2021, a second remote conference was organized focusing on the specific topic we are dealing with. The papers given at this conference have been uploaded onto the project website[5]. The follow-up took the form of a morning workshop where those taking part, from several Italian and French universities, presented original research carried out on the didactics of foreign languages with regard to AI.

The discussion which ensued was so rich that it was decided to select some of the papers from the 23 and 24 April, to revise them in order to

Rachele Raus, Università di Bologna, rachele.raus@unibo.it

[1] To date this research initiative has involved around one hundred researchers from several EU countries, and, in addition, from Canada and Brazil.

[2] See http://www.jmcoe.unito.it/home (last accessed for all sites quoted 20 October 2021).

[3] https://www.dcps.unito.it/do/home.pl/View?doc=toeu.html

[4] https://media.unito.it/?content=9806

[5] http://www.jmcoe.unito.it/

take into account all the meetings of the various working groups and people concerned and to add the contributions of those colleagues who had not been able to take part in the event, thereby further enriching an already plentiful debate. The result is a volume of original unpublished research being made generally available for the first time, at least in Europe.

1. How the issue emerged

As from the 1990s, several international organizations, UNESCO in particular[6], started to question the 'digital fracture', caused, amongst other factors, by certain technologies being more widely available in some countries rather than others. Other similar dangers have emerged, for example the fact that languages that are spoken rather than written are less likely to be represented on the web, that large-scale corpora and big data are increasingly the property of private enterprise[7], that some countries such as Canada or, in Europe, France and Germany, have invested massively in producing pages for the web and more generally in language industries, that most of the linguistic data and corpora are usually available only in English; that multilingual computer tools often use English as the pivot language (see Vetere's chapter in the present issue), etc.

Working groups[8] which we have organized around these questions indicate the following avenues of inquiry:

1. The question of language rights in relation to the possibility some countries have to invest in the language industries compared to some others. This question, raised by both linguists and lawyers in the working groups, was focused on by the United Nations special Rapporteur on Minority Issues, Fernand de Varennes[9], at the 24 April 2021 conference. The debate will continue on the occasion of a conference organized by the Catholic University of Lille in the spring of 2022.
2. The need for careful selection of sources and data used to train AI algorithms.

[6] See http://www.unesco.org/new/index.php?id=50219

[7] Concerning document management for AI, see computer science specialists' talk at the conference on 6 October 2020 at https://media.unito.it/?content=9806

[8] The working groups were divided up by common language (FR, IT, EN).

[9] https://drive.google.com/file/d/1YPSFX-v9ITB4N1tj_eFOl7Sinc9NBgcY/view

This second question raised other closely related issues, first of all the relations that these data have with forms of discrimination, termed 'algorithmic discrimination' by Bartoletti (2020) and that the relationship between these data and sources of 'authority' (*i.e.* authoritative source), which was raised in discussions with Laurent Romany during the 6 October conference. The present situation is that the authoritative source criterion proposed by ISO (2009 : 7) recommends that this source should be an official one or that it should guarantee criteria of quality, which indirectly implies that an institutional source is indeed required. This is probably one of the reasons that all the first training sessions for neural algorithms used in machine translation software or digital devices were performed with multilingual corpora sourced from international organisations. Another is that these institutions have a large quantity of multilingual data which are already available and predigested for automatic processing. That is what happened to the Google translator[10]. Its neural deep learning networks were trained on United Nations corpora (UN 2017: 120), and Facebook used multilingual documents from the European Parliament to train its automatic translators (Le Cun 2019: 277), or the Linguee concordancer[11] — its corpora constitute the training material for the automatic translator DeepL[12], a corpus which comes from EU institutions, UN, the Canadian federal government… That being said, using highly varied corpora and data which are not from institutions and therefore not necessarily checked and validated for quality has the effect of improving the overall performance of IT tools, compared with others using institutional corpora.

One topic that came up in the October conference is the case of the agreement of nouns and adjectives in the feminine. This tends to become particularly problematical in the case of nouns referring to occupations or professions when texts are drafted or translated automatically using tools based on deep learning neural networks trained on institutional corpora, which in turn rely on legal and/or political texts. The 'neutral' masculine then becomes the default option to refer to categories of people. This produces errors in the agreement in the feminine when texts are automatically translated into Romance languages, as is the case for Google or DeepL

[10] https://translate.google.com
[11] https://www.linguee.com
[12] https://www.deepl.com/en/translator

translators, the most frequently used tools by Italian academics (see the contributions in the second part of the issue).

This problem could be solved, at least partially[13], by using different corpora which are not necessarily validated and thus not institutional but which should be closer to actual usage and everyday language. The discussion about the data and the criteria used to establish them turned to the possibility of providing criteria before training neural networks. Thus the human intervention would be transferred from the current practice of revising the texts produced by the machine (i.e. after translation) to a preliminary step, where the linguistic and discursive implementation is used to train AI — selecting specific sub-corpora for specific tasks, providing criteria to improve supervised machine learning, etc.

These criteria would facilitate implementing the recommendations in the 2020 European white paper on human intervention in designing a tool to eliminate linguistic and gender discrimination and to impose 'operational constraints' in favour of multilingualism and linguistic variation (European Commission 2020: 19 and 21).

This is probably a time-consuming solution, that requires previous planning on what the 'authority' is to do in relation to AI training, also involving the task of tagging a large corpus to correspond to the needs of learning. That being said, it may be a time-consuming task but an indispensable one if AI is to fulfil its promise of promoting multilingualism. It is not enough for a machine translation unit to be multilingual if for example it uses English as the pivot language, as is the case for most of the tools produced by the current language industries.

In addition to the issue of language rights and data, the working groups on multilingualism and AI focused on the pedagogical implications, in particular how work in the classroom on languages and AI can be used for learning activities with tools of machine translation based on deep learning neural networks (see the second part of this issue).

[13] Although there are gaps in feminine forms for certain professions and occupations in several Romance languages, notably French and Italian, this tendency is exacerbated in institutional corpora used in training algorithms which regularly resort to the masculine as a 'neutral' form to refer to persons in legal and thence to administrative discourse (see also Marzi 2021).

2. From a critical appraisal to AI promoting multilingualism

Although some contributors may be critical of the consequences of AI on multilingualism (see in particular the articles by Agresti and de Vecchi in this issue) and on a market, such as the language industries market, which today gives a biased view on language as mere vehicles of communication, or even interchangeable 'codes', it is by challenging AI that various ways of using these resources differently in favour of multilingualism can be achieved.

This was an invitation to view the data in a fresh light and to look into ways and means of designing and producing a multilingual application which would facilitate inclusive communication aimed at public service administrations. This application would be designed initially for Italian and then for other European languages (French, then German and finally Spanish and/or English, taking diatopic variation of these languages into account). It would be made with the support of the Polytechnical University of Turin based on criteria designed to develop AI promoting language variation and multilingualism. It is hoped to present the Italian application during the workshop to be organized at the end of 2023, where other similar initiatives will also be on show, such as the new automatic Facebook translator inaugurated in 2020, which no longer uses English as the pivot language[14], or Microsoft AI Azure transcription used by the European Parliament, which takes into account different regional or national accents (Aeles 2020).

This workshop is intended to demonstrate how AI performance can be enhanced if only the right discursive and linguistic choices are made to take on board language variation, be it diaphasic, diatopic or diastratic.

As for the learning section, the working groups were set up during the 2020-2021 academic year and the results of the surveys are presented in the second part of this issue. These surveys have brought out how a specific training strategy can be used to heighten critical awareness of AI, which in turn results in a heightened awareness of the use of AI-based Language Industry tools.

This critical approach in no way undermines the initial confidence that the students say they have in AI, on the contrary it tends to reinforce confidence thanks to a better understanding of how these tools work and what can be made of them.

[14] The development team is coordinated by Prof. Tania Cerquitelli, whom we warmly thank.

In the long run, this issue of *De Europa* aims at promoting the improvement of AI on the one hand by implementing precise linguistic and discursive criteria before the learning takes place and on the other hand by increasing awareness of how certain criteria are favoured in the output. This way, the articles in this volume also intend to contribute to the final research outcomes of the project, planned for 2023, in particular:

- Recommendations for European decision makers on developing policies that favour artificial intelligence which respects multilingualism;
- Guidelines for those who work in vocational areas (technology, IT, standardization organizations, ...), which include suggestions for designing AI algorithms to promote multilingualism and diatopic and diastratic variation of languages, starting with the official languages of the EU.

3. Presentation of the contributions

This issue is made up of articles in one of three languages, Italian, French and English. They present the most salient features of the research carried out by the working groups over the first year and together make up the first volume of a trilogy, each volume presenting the results of the year's work. The approach is a very broad one so that highly specialized information can reach as many people as possible without compromising the scientific approach. A deliberate policy has been to use (gender)inclusive formulations, though no pressure has been put on any of the contributors to adopt any specific language choice. This results in some chapters preferring feminization, using the median point, hyphens or brackets, and other chapters — the majority in fact — where the choice has been for the so-called 'neutral' masculine.

This issue is in two parts. The first section is a general overview on AI as it is related to multilingualism and language variation. It also contains a number of specific case studies in particular concerning machine translation and AI, which go to make up the last part of this section. The second part of the volume contains the results of various surveys made in a teaching and learning context carried out in a number of European universities and business schools which took part in the initiative, namely Kedge Business School, EDEH, EM-Normandie, the Montpellier Business

School and the École supérieure de commerce de Clermont (Clermont Business School), the Catholic University of Lille, the Catholic University (Sacro Cuore) of Milan, the University of Naples-Orientale, the University of Turin and the University of Verona. This is also where some articles and studies are carried out by university students (first or second year Masters or postgraduate students), since the project was conceived as a research-action.

Before inviting our readers to discover the contents of this special issue we would like to thank John Humbley, Alida Maria Silletti and Silvia Domenica Zollo for their tireless efforts and punctual revision and formatting of the articles and for their participation in all the thought which has gone into the project which has come to fruition. We would also like to thank Umberto Morelli, who coordinates the project *Artificial Intelligence for European Integration* and who has always managed to steer the working groups in the right direction, even when the conditions became difficult notably with the Covid-19 pandemic.

We must also thank all those who contributed directly or less directly in the debates on the various topics.

We hope that the debate will continue and become richer as it continues, since artificial intelligence has become an indispensable part of our daily lives.

References

Aeles Joris (2020). *Rapprocher le Parlement européen des citoyens grâce à l'IA*. https://pulse.microsoft.com/fr-be/transform-fr-be/government-fr-be/fa1-rapprocher-le-parlement-europeen-des-citoyens-grace-a-lia/

Bartoletti Ivana (2020). *An Artificial Revolution. On Power, Politics and AI*. Edinburgh: Indigo.

European Commission (2020). *White paper on Artificial Intelligence — A European approach to excellence and trust*. COM(2020) 65 final.

International Organization for Standardization — ISO (2009). *Standard 23185. Assessment and benchmarking of terminological resources*. Geneva: ISO.

Le Cun Yann (2019). *Quand la machine apprend*. Paris: Odile Jacob.

Marzi Eleonora (2021). "La traduction automatique neuronale et le biais de genre : le cas des noms de métiers entre l'italien et le français". *Synergies Italie*, 17, 19-36. https://gerflint.fr/Base/Italie17/marzi.pdf

United Nations (2017). *Change. Rapport annuel 2017*. Geneva: United Nations.

Première partie : réflexions et études de cas

Introduction
Réflexions et études de cas à l'aune de l'intelligence artificielle. Vers de nouveaux observables linguistiques ?

John Humbley, Silvia Domenica Zollo

L'intelligence artificielle (IA) représente l'un des sujets d'ébranlement majeurs qui affectent notre époque. Très rarement une évolution technologique n'aura engendré autant d'occasions de résolutions de problèmes et autant de transformations dans la société et dans la recherche scientifique. Des grandes bibliothèques aux salles de classe, des laboratoires aux entreprises, aucun échelon de nos structures ne semble pouvoir échapper à la révolution des réseaux neuronaux artificiels, aux architectures profondes, aux couches cachées et au *deep learning*. Ce type de bouleversement n'est pas nouveau, surtout si l'on pense au père de l'ordinateur et de l'IA, Alan Turing, qui aurait aujourd'hui plus de 100 ans. Toutefois, la capacité de stocker un nombre indicible de données sur une base numérique et la puissance actuelle de l'automatisation des traitements de données rendent l'IA désormais indispensable.

Éclairer autant que possible les mécanismes de l'IA et en expliquer la valeur ajoutée en matière de multilinguisme, variétés linguistiques et traduction, tel est le but de cet ouvrage qui réunit linguistes, terminologues et informaticiens dans un dialogue interdisciplinaire et authentique autour de ces thématiques. La première partie de ce recueil — *Réflexions et études de cas* — veut montrer les possibilités nouvelles que l'IA offre aux chercheur·e·s en terminologie, traduction et analyse de corpus, en donnant à voir des représentations originales, en objectivant des parcours de lecture heuristiques et en faisant émerger de nouveaux observables linguistiques. Le point commun des contributions qui vont suivre est d'examiner de manière critique et concrète les avantages et les inconvénients de l'IA appliquée aux études linguistiques et de proposer un pro-

John Humbley, Université de Paris Cité, humbley.john@orange.fr
Silvia Domenica Zollo, Università di Verona, silviadomenica.zollo@univr.it

gramme de recherche dont nous souhaitons ici poser les premières bases. Cette partie se compose de sept contributions que nous avons classées en trois sous-sections, selon la pluralité des points de vue sur l'IA et la diversité des approches — épistémologique, linguistique et informatique.

La première sous-section est consacrée à l'apport — actuel ou potentiel — de l'IA aux langues dites minoritaires en Europe et plus généralement à la promotion de la diversité linguistique et du multilinguisme. C'est Giovanni Agresti qui ouvre la réflexion en proposant plusieurs pistes complémentaires pour bien comprendre la question préalable mais souvent mal posée de minorité linguistique. Dans *Intelligence artificielle et langues minoritaires : du bon ménage ? Quelques pistes de réflexion*, il insiste sur les différentes interprétations qu'il convient d'associer à ce terme selon le contexte : c'est ainsi que toute langue autre que l'anglais peut être minoritaire, surtout dans le cadre de l'Union européenne et que de nombreuses langues européennes doivent être qualifiées d'ultra-minoritaires. Comme le disait Mackey « Ce n'est pas le nombre seul de locuteurs, quel qu'il soit, qui fait une minorité, c'est l'ensemble des fonctions de sa langue » (1996 : 279).

C'est en examinant les processus de minorisation linguistique[1] que l'on arrive à mieux cerner les chances de survie et de développement de ces langues menacées de disparition, avec ou sans l'aide de l'IA selon la situation. L'essor de l'Internet a ouvert des espaces d'échange dans des communautés linguistiques où la langue n'est plus guère parlée, permettant aux langues concernées de remplir des fonctions nouvelles. Il en ressort que l'IA, qui part du principe d'une langue fortement standardisée et qui requiert de gros corpus, n'est pas nécessairement adaptée à ces langues, surtout aux ultra-minoritaires. L'auteur entrevoit toutefois des possibilités de permettre de « retrousser la diglossie » grâce à l'IA, par la traduction automatique (TA) par exemple, de telle sorte que les représentations sociales de ces langues se trouvent améliorées. Ce recours à l'IA reste néanmoins une possibilité et non une nécessité : le plus important, c'est que ces langues possèdent encore une fonction, qui peut être identitaire, pour qu'elles perdurent.

Dans son article intitulé *Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive*, Guido Vetere entre dans le vif du sujet en abordant de front les questions techniques comme

[1] Au sujet de la minorisation linguistique, on signale la publication récente de Giovanni Agresti (2021).

préalable d'un traitement linguistique. Il s'avère que toute langue autre que l'anglais est désavantagée par rapport à l'accès aux défis de l'automatisation de la langue. L'auteur commence par faire un état des lieux des technologies de la langue par rapport aux réseaux neuronaux avant de considérer comment les adapter en vue d'applications dans d'autres langues. Il entame ensuite un passage en revue critique des projets et des politiques européens visant à rééquilibrer la situation linguistique. La dernière partie de l'article est un plaidoyer pour une approche lexicale en exploitant des ressources comme Wordnet, plutôt que statistiques pour alimenter les réseaux neuronaux.

La réflexion sur les langues minoritaires européennes se poursuit dans la deuxième sous-section avec l'étude de Kepa Sarasola, d'Itziar Aldabe et de Nora Aranberri, *Enabling additional official languages in the EU for 2025 with language-centred AI*, qui examinent le cas du basque, langue que Giovanni Agresti qualifie de relativement bien équipée car ayant des institutions dynamiques de normalisation et disposant de corpus relativement importants, tout en étant handicapée en tant qu'isolat linguistique. Les auteur·e·s se focalisent sur la promotion des langues comme le basque au niveau européen, rendue désormais possible grâce à l'aide de l'IA. En effet, les auteur·e·s commencent par exposer les raisons qui justifient cette recherche de statut plus officiel au niveau européen avant de proposer des mesures qui deviennent de plus en plus réalisables grâce au recours aux outils comme la TA ou la reconnaissance et la synthèse vocales. Les technologies de la langue, dont plusieurs sont tributaires de l'IA, sont thématisées dans l'optique de la revitalisation de langues comme le basque qui profitent d'un nouvel élan de mesures en leur faveur. Les auteur·e·s décrivent en détail les différents outils désormais susceptibles de rendre des services comme pour la traduction de Wikipédia ou de Wikidata ou pour des guides multilingues de musées, et mettent en avant les avantages des quelques projets européens destinés aux langues manquant de ressources. Toutes ces avancées permettent d'envisager la promotion de certaines langues comme le basque à un statut plus officiel au sein de l'Union européenne, ce que les auteur·e·s appellent les langues de genre tiers.

Tout en restant dans le domaine du multilinguisme, une deuxième question est posée à propos du récent dialogue entre IA, diversité linguistique et connaissances spécialisées, thème approfondi par Maria Luisa Villa, Maria Teresa Zanola et Klara Dankova dans *Langages et savoirs : in-*

telligence artificielle et traduction automatique dans la communication scientifique. L'attention est accordée au rôle de la TA pour la mise en œuvre d'une communication scientifique valide permettant de préserver la diversité linguistique et culturelle transmise par les langues et les terminologies spécialisées. Les auteures débutent par une introduction sur le rôle de l'anglais, qui s'est affirmé au niveau mondial en tant que langue de la communication scientifique et technique[2]. Si, d'un côté, ce rôle de *lingua franca* a le mérite de favoriser les échanges dans un contexte scientifique de plus en plus internationalisé, de l'autre, cette hégémonie linguistique limite la transmission des connaissances scientifiques au sein des sociétés des pays non anglophones. D'autre part, une culture scientifique majoritairement véhiculée par la langue anglaise génère des inégalités et n'encourage pas la démocratisation de l'accès au savoir produit par la recherche, l'un des principes clés de la science ouverte. Dans le cadre de la crise sanitaire liée à la pandémie de Covid-19, qui a fait apparaître la nécessité de diffuser les connaissances scientifiques et médicales auprès des citoyens, de nombreuses recherches ont mis en lumière ce phénomène d'anglicisation dans la communication scientifique, qui risque de ne pas répondre pleinement à la troisième mission de la science, à savoir informer les citoyens dans leurs langues maternelles (Taşkın *et alii* 2020). Dans ce contexte, comme signalé par les auteures, plusieurs initiatives ont vu le jour depuis 2020, afin de rappeler la primauté du multilinguisme dans la communication scientifique. Parmi celles-ci, le lancement de la revue *Nature Italy*, un supplément en ligne de la revue *Nature*, dont l'un des premiers numéros est consacré à la lutte contre la pandémie de Covid-19 dans les laboratoires de recherche italiens.

Afin de pallier ces difficultés et de favoriser le multilinguisme à une large échelle dans la communication scientifique, les auteures insistent sur la possibilité d'explorer les nouveaux chemins ouverts par l'IA, tout particulièrement de recourir à des technologies de la TA de plus en plus performantes. En ce sens, elles identifient des bonnes pratiques d'usage dans le but de briser la barrière linguistique représentée par l'usage dominant de l'anglais pour la diffusion des connaissances scientifiques, et des pistes d'action didactique afin d'initier les étudiant·e·s à la traduction de la production

[1] Depuis les années 1960, par exemple, la langue française souffre d'un déficit lexical de milliers de mots chaque année au regard de la progression enregistrée pour l'anglais (Bowker *et alii* 2019).

scientifique, en s'appuyant sur les outils de TA et sur un corpus d'articles publiés, entre autres, dans des revues de vulgarisation scientifique.

La troisième sous-section est dédiée à un approfondissement de l'apport de l'IA à la terminologie et à la TA. Elle s'ouvre tout naturellement sur la contribution d'Anne Condamines, *Terminologie, intelligence artificielle, psychologie cognitive : réflexions sur les interactions possibles dans l'étude de la variation en langue spécialisée*, qui fait le point sur les progrès accomplis en TALN (Traitement Automatique du Langage Naturel) et en terminologie depuis une trentaine d'années, et qui redéfinit un certain nombre de méthodes et d'objectifs de recherche en relation avec l'IA.

L'auteure propose, tout d'abord, une série d'interrogations concernant les possibilités de rencontre entre terminologie, IA et psychologie cognitive, ce qui constituera le fil conducteur de toute sa contribution. À son avis, la terminologie est à un moment crucial de son histoire ; le rapprochement avec l'IA et la dimension « émotionnelle », qui fait écho aux théories de la psychologie cognitive, devraient l'aider à mieux structurer et organiser les connaissances dans la communication technique et à répondre aux besoins des analystes de la variation dans les corpus spécialisés.

Après avoir présenté des éléments de réflexion susceptibles de contribuer à l'avancement de la recherche sur les interactions avec le TALN et l'IA, elle propose deux études : la première sur la présence/absence de prépositions dans la construction de *pêcher* et *rivière* en français et *to fish* et *river* en anglais ; la seconde sur le contexte lexical de chaque structure en anglais. Les résultats montrent que les constructions syntaxiques se différencient selon plusieurs critères (le genre textuel, le rapport subjectif et affectif des locuteurs, dans ce cas les pêcheurs, avec la rivière) et que l'environnement lexical fait partie de catégories sémantiques différentes. L'étude se conclut par une évaluation sur les retombées positives et négatives de l'IA dans des recherches qui visent à considérer les facteurs extralinguistiques dans la description des phénomènes de variation sur corpus.

Les deux dernières contributions thématisent la place de l'IA dans les études traductologiques. C'est ainsi que Christopher Gledhill et Maria Zimina explorent l'un des principaux impacts des dernières avancées en matière de traduction automatique neuronale (TAN) sur un cours de traduction de Master à l'Université de Paris, à savoir la nécessité d'enseigner aux futurs traducteurs et aux futures traductrices des compétences de révision avancées sur la base des résultats de la TAN. Dans *Human-machine*

interaction: how to integrate plain language rules in the revision cycles of Neural Machine Translation output, les auteur·e·s affirment qu'un des éléments cruciaux de ces compétences est la capacité de répondre aux exigences de qualité des utilisateurs en matière de langue normalisée, telle que la « langue claire » (LC), désormais requise par les politiques de communication et les guides de style de nombreuses organisations. Pourtant, les principes de la LC ne sont pas toujours bien définis, ni mis en œuvre de manière uniforme dans les textes sources. Par conséquent, les résultats de la TAN peuvent présenter des lacunes en termes de LC, ce qui nécessite de nombreux cycles de post-édition (PE) et de révision. Pour l'apprentissage de la traduction, qui prévoit également l'apprentissage de la révision et de l'édition des textes en sa langue maternelle, la maîtrise des principes de la LC peut effectivement constituer une difficulté supplémentaire.

Après avoir présenté une expérience d'intégration des principes de la LC dans le cours de Master 2 en Traduction de sites web vers l'anglais, les auteur·e·s examinent les retombées de la TAN dans les phases de PE et de révision en termes de personnalisation et de correspondance des mémoires de traduction. Enfin, ils concluent en suggérant l'intégration de la LC — souvent négligée dans les débats actuels sur l'impact de la TAN — qui pourrait apporter une valeur ajoutée aux projets de traduction et permettrait aussi de défendre la notion de service de qualité traductive.

Qui dit TAN dit généralement progrès, un aspect qui est bien reflété dans la dernière contribution, *A Journey in Neural Machine Translation*, signée de Philippe Langlais, soucieux de provoquer une réflexion théorique sur les dernières évolutions en matière de TA, grâce aux progrès de la TAN.

Il est désormais admis que la TAN produit des traductions plus fluides avec beaucoup moins d'erreurs que les systèmes de traduction automatique statistique (TAS) précédents. Toutefois, plusieurs avis différents coexistent sur ce point. Si d'un côté, pas mal d'études (Toral *et alii* 2018) ont largement démontré que la TAN permet de travailler plus rapidement et qu'elle n'a aucun impact négatif sur la qualité traductive, de l'autre, un bon pourcentage de chercheur·e·s a fait état des problèmes liés à la TAN, tels que le gain de temps et la qualité de la PE sortie des moteurs de TAN (Castilho *et alii* 2017). D'où, selon l'auteur, l'urgence de mettre en place de nouvelles recherches pour vérifier l'impact des moteurs de TAN les plus récents, les effets des systèmes neuronaux qui intègrent une mémoire de traduction et ceux des systèmes au niveau des documents.

Langlais est convaincu que la TAN actuelle, utilisée correctement, peut représenter un véritable atout dans les contextes professionnels. Dans sa contribution, il attire l'attention sur le fait que la TAN n'est pas une affaire facile : elle requiert une expertise assez spécifique et des outils informatiques appropriés. Sans cela, une organisation souhaitant utiliser la TAN n'a d'autre choix que de traiter directement avec des prestataires de services de traduction, qui se chargeront d'adapter la technologie aux spécificités des besoins. Il discute ainsi du type d'expertise nécessaire à l'entraînement d'un moteur neuronal, à l'aide de librairies SOTA (State-of-the-art). Ensuite, il s'intéresse à la qualité produite par les moteurs neuronaux et montre que les données sont un élément important dans le développement d'une solution de TAN solide. L'auteur conclut en abordant d'autres options d'intégration de la TAN qui pourraient s'avérer plus utiles que la simple PE.

La très grande variété de ces contributions a, nous semble-t-il, permis d'amorcer un dialogue entre théoriciens et expérimentalistes, suscitant ainsi de nouvelles idées sur la relation encore ambivalente que l'IA entretient au système linguistique. Les résultats et les retours d'expérience des auteur·e·s semblent suggérer que l'IA peut produire des résultats positifs pour les linguistes et qu'une nouvelle culture de travail est en train d'émerger. Il s'agit d'une culture annonciatrice d'une attitude optimiste, voire enthousiaste, par rapport à l'IA. Nous sommes face à une nouvelle révolution du langage humain, et comme pour toute révolution, ce qui pour certains peut être un succès et un motif de satisfaction peut susciter chez d'autres la prudence, voire la crainte. À ce propos, un aspect que l'on pourrait éventuellement prendre en considération dans les recherches futures est la nécessité d'appliquer des normes juridiques et éthiques pour garantir un accès équitable et inclusif à l'information dans toutes les langues. Avant de déployer le *deep learning* à grande échelle, il serait bon de lui inculquer quelques règles « morales » et d'entraîner la machine à un système d'apprentissage avec des prescriptions qu'elle serait contrainte de respecter en toute langue. Bien évidemment, l'ambition de parvenir à imiter une cognition humaine éthique et consciencieuse au préalable nécessiterait de nouvelles découvertes interdisciplinaires en recherche fondamentale et non une simple évolution des technologies actuelles d'apprentissage automatique. Sur ce point, le dialogue reste donc encore ouvert.

Concernant les études en terminologie et en traduction multilingue, il est évident que l'IA peut certainement jouer un rôle essentiel dans le monde de la recherche, de plus en plus internationalisé. Bien sûr, les outils de TA et de gestion terminologique ont un coût très élevé pour les institutions de recherche, mais il est envisageable d'optimiser l'organisation des données multilingues à l'aide de technologies de plus en plus performantes, surtout pour élargir l'accès à l'information technique et scientifique à plusieurs couches de la société et en toute langue, notamment dans les langues minoritaires.

Bibliographie

Agresti Giovanni (2021). "Bien nommer pour bien agir ? La notion de '*minoranza linguistica*'" (minorité linguistique) en Italie et la genèse des lois et politiques linguistiques". In : Alain Viaut (ed). *Catégories référentes des langues minoritaires en Europe*, Pessac : Maison des Sciences de l'Homme d'Aquitaine, 285-296.

Bowker Lynne, Buitrago Ciro Jairo (eds) (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. UK : Emerald.

Castilho Sheila, Moorkens Joss, Gaspari Federico, Calixto Iacer, Tinsley John, Way Andy (2017). "Is Neural Machine Translation the New State of the Art?". *The Prague Bulletin of Mathematical Linguistics*, 108, 109-120.

Mackey William (1996). « Langue première et langue seconde ». In : Hans Goebl, Peter H. Nelde, Zdeněk Starý, Wolfgang Wölck (eds). *Kontaktlinguistik/contact linguistics/ Linguistique de contact*, Berlin : De Gruyter, 271-283.

Taşkın Zehra, Doğan Güleda, Kulczycki Emanuel, Zuccala Alesia Ann (2020). "Long read. Science needs to inform the public. That can't be done solely in English". In: LSE Covid 19 Blog, 18 june 2020, URL: http://eprints.lse.ac.uk/105125/

Toral Antonio, Wieling Martijn, Way Andy (2018). "Post-editing Effort of a Novel with Statistical and Neural Machine Translation". *Frontiers in Digital Humanities,* 5, 2297-2668.

Quelques réflexions sur le multilinguisme à l'aune de l'intelligence artificielle

De Europa Première partie - Special Issue 2022

Intelligence artificielle et langues minoritaires : du bon ménage ? Quelques pistes de réflexion

Giovanni Agresti

« Es sus la talvèra qu'es la libertat »
Joan Bodon / Jean Boudou

Introduction

L'intelligence artificielle (IA) est un horizon technologique poursuivi depuis longtemps, comme le témoignent quelques mythes et quelques œuvres littéraires (le golem, Frankenstein, l'Ève future, etc.). On peut s'interroger quant à la nature de cette instance : volonté de puissance divine ? Désir (masculin) de génération par la science et la technique ? Outillage et prolongement de l'intelligence humaine ? Quoi qu'il en soit, l'essor du numérique, et tout particulièrement la capacité d'« apprendre à apprendre »[1] d'ordinateurs à la puissance de calcul toujours plus élevée, semble justifier ce syntagme (« intelligence artificielle »), qui garde néanmoins l'allure d'un oxymore chez nombre d'auteurs et dont la légitimité est parfois foncièrement remise en question (Julia 2020).

Nous ne sommes guère des spécialistes de la matière. Pourtant, nos pratiques professionnelles sont de plus en plus confrontées à l'IA, notamment dans le domaine du traitement numérique du langage. Plus en général, nous nous interrogeons sur l'intérêt, voire l'engouement que suscite ce thème, tellement porteur qu'il semble déjà impensable de ne pas considérer l'IA comme l'horizon ultime de notre intelligence, comme si paradoxalement le progrès de l'espèce ne pouvait passer que par une croissante autonomie de la machine. Autrement dit, l'IA est en train de devenir un impensé, une évidence, une fatalité, quelque chose d'inéluctable et de nécessaire — la solution, sans doute, à la plupart de nos problèmes ? Ce

Giovanni Agresti, Université Bordeaux Montaigne, giovanni.agresti@u-bordeaux-montaigne.fr
[1] Cette formule ne manque pas d'ambiguïté : il ne s'agit pas de l'apprentissage réflexif prôné par des auteurs tels que, entre autres, Edgar Morin (1986) ou Idries Shah (2009).

constat serait déjà suffisant à mobiliser notre analyse, l'une des missions du chercheur, notamment en sciences humaines et sociales, étant précisément de « ne pas » aller dans la direction du courant majoritaire, qui avance tout seul et qui n'a pas besoin de laquais déguisés en docteurs de recherche…

Cela dit, la raison principale qui justifie notre expertise concerne les langues naturelles, notamment les langues dites « minoritaires », et leur traitement — diffusion, promotion, occultation, etc. — par l'IA en particulier et par le numérique en général. Ce dernier étant aussi bien un champ de communication qu'un terrain de pouvoir, notre réflexion concerne par ricochet également le dossier des droits et politiques linguistiques. Dans les paragraphes qui suivent nous proposons juste quelques pistes de réflexion visant à complexifier le thème du rapport entre l'IA et les langues minoritaires, rapport qui ne va pas de soi et qui, disons-le tout de suite, n'a rien d'évident. Nous commençons par nous interroger sur la notion de « langue minoritaire », beaucoup plus problématique qu'il n'y paraît de prime abord (Par. 1) ; de fil en aiguille, nous questionnons la dimension virtuelle des langues à l'ère de l'Internet (Par. 2) ; pour conclure, sur ces bases théoriques nous tâchons de vérifier quel rapport peut-il exister entre l'IA et la documentation, la pratique et le développement des langues minoritaires. Deux questions majeures se posent dès lors : le rôle des *big data* et de la normativisation ; l'exploitation de l'IA dans la perspective de la revitalisation de langues moins répandues (Par. 3). Un bilan conclusif va clôturer notre réflexion tout en suggérant quelques ouvertures de perspective.

1. La notion de « langue minoritaire » : ni une, ni indivisible

Si nous ne sommes pas des spécialistes de l'IA, nous fréquentons depuis longtemps les langues minoritaires et le problème connexe de leur catégorisation, c'est-à-dire leur typologie sociolinguistique, notamment en contexte européen[2]. Plusieurs enseignements découlent de cette expérience prolongée et nourrie du terrain :

[2] Nous signalons à ce sujet la base de données textuelle en libre accès CLME (Catégorisation des Langues Minoritaires en Europe), conçue par Alain Viaut et actuellement hébergée dans le site de la MSH de Bordeaux : https://www.mshbx.fr/base-clme/. Dernière consultation : 25 octobre 2022.

a) La catégorie « langue minoritaire » est porteuse de malentendus à n'en plus finir. *A minima*, il faut d'abord distinguer entre des langues qui, tout en étant minoritaires à l'échelle nationale ou continentale, ne le sont pas tout à fait à l'échelle régionale ; et des langues qui, en revanche, sont toujours et partout minoritaires. Classiquement, on convoque, pour le contexte européen, la langue catalane, minoritaire en Espagne mais tout de même pratiquée par quelque 7 à 10 millions de locuteurs, suivant le type de statistique, l'inclusion ou pas du catalan de Valencia dans ce domaine linguistique, la nature et le niveau de maîtrise linguistique des locuteurs pris en compte (incluant ou pas les semi- et les néo-locuteurs), etc. Il est évident que cette langue n'a pas de statut comparable à celui d'une langue locale très peu diffusée et peu vitale, par exemple le grec de Calabre qui n'est plus pratiqué que par quelques dizaines de personnes, qui plus est très âgées. D'où la nécessité, au préalable, d'articuler notre ensemble au moins en deux sous-ensembles : « langues minoritaires » et « langues ultra-minoritaires »[3].

b) Cette première, sommaire distinction, qui privilégie l'aspect purement quantitatif, présente de fortes retombées également du point de vue qualitatif. Une langue pratiquée par plusieurs millions de personnes possède une valeur économique considérable, et donc un plus fort attrait, surtout si cette langue a le statut de langue co-officielle et si ses locuteurs partagent un territoire suffisamment dense et défini. En effet, dans ces conditions une langue minoritaire régionale a plus de chances de résister à l'érosion, de même qu'une grosse masse de neige fond d'autant plus lentement si elle est concentrée et si sa surface exposée au soleil est réduite. À l'opposé, une langue fortement dispersée, par exemple le rromani, est beaucoup plus menacée de pulvérisation et d'étiolement. À côté des « langues minoritaires » il faut donc parler aussi de processus et de conditions de « minorisation ».

c) Dans l'évaluation du risque de minorisation d'une langue, le niveau typologique entre également en ligne de compte, pourvu que l'on se mé-

[3] À ce sujet, le juriste Giovanni Poggeschi (2012) propose la catégorisation « *iperminoranze* » (« hyper-minorités ») pour indiquer ces groupes qui, indépendamment du contexte et de la nature de leur assise territoriale, sont toujours et partout minoritaires, marginalisés. Dans l'économie de notre exposé, par souci de respect de la consigne (voir *infra*, note 13) nous privilégions la composante linguistique (langues minoritaires) au lieu de la composante ethnique (minorités linguistiques).

fie de toute tentation déterministe. Une langue minoritaire proche (typologiquement et géographiquement) d'une langue de large diffusion peut être « absorbée » par celle-ci, rien que par le fait d'être considérée, à tort ou à raison, l'un de ses dialectes. Ainsi, substantiellement tous les parlers régionaux d'Italie sont pensés par la doxa comme étant des dialectes de la langue nationale, l'italien standardisé à partir de la variété toscane, et ce indépendamment de la distance étique qui sépare ces variétés, parfois très significative. En revanche, une langue typologiquement connotée, par exemple le basque, qui est un isolat linguistique[4], ne saurait être ramenée à la version « dialectale » ou « corrompue » d'aucune autre langue.

d) Ce dernier fait présente des avantages indéniables (impossible de ne pas considérer une variété linguistique « isolée » comme autre chose qu'une langue à part entière) mais également des risques. Précisément parce que langue sans parenté génétique avérée avec aucune autre langue, elle est fatalement très difficile à apprendre : d'où un rapport économique défavorable entre valeur d'usage et coûts d'apprentissage.

e) Cependant, les choses ne sont pas aussi simples que cela : une valeur économique limitée peut correspondre à une valeur symbolique, identitaire, très forte, qui à son tour peut, si la politique suit, se traduire en valeur économique et, plus largement, en valeur territoriale[5]. Finalement, une langue (ultra-)minoritaire peut être vue aussi comme la langue d'une élite.

f) Sous certaines conditions, la proximité entre langues peut également s'avérer un atout : deux langues minoritaires régionales comme le catalan et l'occitan forment un diasystème qui les amplifie ; une langue minoritaire proche d'un ensemble de langues, dont des langues nationales — par exemple l'occitan par rapport à l'espagnol ou l'italien et, plus en général, par rapport à toutes les langues romanes —, peut représenter une sorte de passerelle ou charnière entre celles-ci. Le thème

[4] Rappelons que, en typologie, un « isolat linguistique » (souvent malencontreusement confondu avec « îlot linguistique ») représente une variété linguistique sans parenté génétique prouvée avec aucune autre langue naturelle.

[5] Ces enjeux relèvent en large mesure de la linguistique pour le développement (www.poclande.fr), qui se doit d'explorer les voies qui permettent d'améliorer les conditions d'existence de groupes humains à partir du « traitement » de leur(s) langue(s) et mémoires.

de l'intercompréhension entre langues voisines est de mise aujourd'hui et peut modifier sensiblement les représentations et la valeur économique des langues en jeu.

g) La condition de « langue (ultra-)minoritaire » est précisément une « condition » et très souvent dépend également du contexte d'observation. La langue italienne, nationale et officielle (*de jure* et *de facto*), indispensable pour vivre et évoluer en Italie, est assurément minoritaire dans le contexte des travaux des institutions européennes. La langue française est aujourd'hui très souvent minorisée dans les congrès internationaux, y compris lorsqu'elle est l'une des langues officielles de travail. En une formule, « nous sommes tous minoritaires » (Agresti 2016), à un moment ou un autre, à tel ou tel endroit. Mais il est vrai (voir plus haut, note 3) qu'il existe des langues et des communautés linguistiques qui sont toujours et partout minoritaires, marginales, périphériques.

De ces considérations on peut conclure que le statut de « langue minoritaire » n'est pas donné une fois pour toutes et que, à quelques exceptions près, il est plutôt une condition, variable (et modifiable) dans le temps, l'espace, le contexte. Cette notion doit être questionnée pour ce qui est de sa valeur : est-ce que le caractère « minoritaire » est forcément et toujours une diminution ? Par ailleurs : quel est le rapport entre langue majoritaire/ minoritaire et discours majoritaire/ minoritaire ? Comment et pourquoi une langue devient-elle minoritaire ? Quel est le rôle joué par la technologie numérique dans tout cela ? Dans les paragraphes qui suivent nous tâchons, entre autres et un tant soit peu, de répondre à ces questions. D'ores et déjà nous tenons à préciser un premier élément : on ne peut pas parler du rapport entre l'IA et les « langues minoritaires » en faisant fi de la formidable complexité de cas de figure, de points de vue, de statuts de ces langues.

2. La dimension virtuelle des langues à l'ère d'Internet

On a l'habitude de quantifier, voire catégoriser les langues (majoritaires, minoritaires…) sur la base du nombre de leurs locuteurs. De manière plus neutre, on devrait plutôt dénombrer les pratiquants de la langue. En effet, une langue peut être parlée, certes, et c'est le plus important dans les régimes de transmission et acquisition familiales, dont l'ora-

lité est à la fois le cadre et l'outil naturel et premier ; mais une langue peut être aussi pratiquée autrement, à l'écrit, et parfois même exclusivement, en contexte minoritaire, lorsque la transmission familiale est coupée ou affaiblie. On a pu observer ce phénomène en Corse : l'essor de blogs en langue corse correspond à l'essor d'une pratique écrite de la part des nouvelles générations qui souvent n'ont pas reçu la langue directement de leurs parents ou grands-parents (Quenot 2010). Cet exemple suggère une autre considération : l'Internet représente un espace virtuel qui est aussi un espace de communication et linguistique alternatif, que les jeunes s'approprient plus facilement. D'où une tendance à un changement très significatif des pratiques linguistiques à l'ère et à cause de l'Internet.

D'une manière très générale, depuis que l'Internet s'est affirmé, pleinement normalisé dans la société, la question de savoir dans quel sens et dans quelle mesure cet espace « autre » (ou *cyberespace*) pouvait doubler, compliquer, remplacer, étoffer, etc. l'espace réel n'a eu de cesse de se poser. Concernant les langues naturelles, cette question a pris souvent la forme d'une mise à jour de la formule à succès — à tout le moins fort discutable — de la « guerre des langues » (Calvet 1999a). À tour de rôle terrain de conquête économique, lieu de démocratie, champ d'exercice du pouvoir, etc., la Toile est dans tous les cas un vaste, immense territoire où la communication suit des « lois » qui ne sont pas toujours les mêmes que dans le monde réel — même si, à bien y voir, désormais ce monde virtuel est tellement envahissant qu'il doit être considéré comme pleinement intégré dans la vie de tous les jours. La séparation entre réel et virtuel, et partant entre communication réelle et virtuelle, est sans doute aujourd'hui beaucoup plus nuancée et controversée qu'elle ne l'était il y a vingt ou trente ans.

Quoi qu'il en soit, la Toile est une logosphère, à savoir un monde où l'essentiel est le dépôt et le partage d'informations, de discours, d'images et de représentations au sens large. Cette logosphère est analysable, l'écrasante majorité des actes de langage intervenant en absence, c'est-à-dire à l'écrit. On a pu affirmer que jamais on n'a autant écrit qu'à l'époque actuelle — bien entendu pour le meilleur et pour le pire. Or, dans cette logosphère on peut mesurer le « poids » des différentes langues employées. On observe, d'une manière générale, deux tensions opposées : d'une part, une tension centralisatrice : la Toile étant un réseau de réseaux, s'il est vrai qu'elle fait de tout point du système un centre, il n'en demeure pas moins que ce centre, pour rayonner à l'international, se doit de privilégier les

langues « hypercentrales » (Calvet 1999b). D'autre part, la Toile est aussi un espace communautaire, où (re)nouer des liens de proximité (par exemple entre groupes diasporiques ou dispersés), où exister parallèlement à la société « réelle », où se rassembler dans la marge, si besoin. Où, aussi, la norme se fait plus souple. La Toile est donc traversée aussi par une forte tension décentralisatrice, où la diversité linguistique — des langues minoritaires aux langues cryptiques — s'invite volontiers, aussi au vu de la facilité d'y publier des textes. Un dernier élément qui mérite d'être évoqué dans cette réflexion liminaire sur les langues à l'ère de l'Internet est la complexification des registres. La distribution traditionnelle entre oral et écrit n'est plus satisfaisante dès lors que s'affirment, par exemple, l'écriture instantanée et l'oralité asynchrone[6].

3. Langues minoritaires et IA : du bon ménage ?

La numérisation des langues et leur circulation dans la Toile, la complexification des registres et des réseaux, le caractère envahissant de la communication digitale, le régime multilingue et pluriglossique qui gouverne l'Internet et la richesse des articulations que prend la notion de « langue minoritaire » font du numérique en ligne et hors ligne un terrain de pratiques langagières extrêmement diverses et à la lecture difficile, ou à tout le moins peu linéaire, presque insaisissable. Par ailleurs, ce terrain est également un champ économique et donc, par conséquent, un terrain de pouvoir.

Or, là où il y a pouvoir, il y a également des hégémonies et des forces minoritaires plus ou moins assujetties à celles-ci. Ainsi, la valeur économique des langues se reflète *grosso modo* dans la hiérarchie de leurs diffusion et usage en ligne. *Grosso modo*, car la Toile est aussi un lieu public d'émergence démocratique, où se manifestent des formes de contre-pouvoir ainsi que des actes de résistance, souvent surprenants : les gouvernements illibéraux tendent à museler l'Internet, et tout particulièrement les réseaux sociaux, source de parole et de discours autres. Le centre essaie de contrôler la périphérie, et pour ce faire, de tous temps, il a fallu identité de langue. La grande nouveauté, aujourd'hui, est que l'IA permet de surveiller cette communication de manière très efficace, et plus la langue de

[6] Pour aller plus loin dans la réflexion sur le rapport entre les langues minoritaires et le web, cf. Agresti (2017).

communication est hégémonique, plus le contrôle sur les contenus qui circulent dans la Toile s'exerce dans un environnement pervasif. Par conséquent, l'usage en ligne de langues minoritaires, *a fortiori* ultra-minoritaires, pourrait se configurer comme une sorte d'acte de langage en soi « subversif », en ce qu'il échappe plus facilement au contrôle (réel ou juste potentiel) des maîtres de la communication digitale. Dans ce cas de figure, IA et langues minoritaires ne font pas bon ménage.

Bien évidemment, tout n'est pas contrôle et punition (Foucault 1975) dans la vie de la cité numérique, y compris à l'ère de l'IA ! La périphérie (linguistique) peut tout à fait tirer profit de la tension centralisatrice de la Toile, comme l'a démontré, il y a quelques années, la traduction en langue frioulane du site web de l'Udinese, équipe de football de première division italienne (Serie A) de la ville d'Udine et la chronique en frioulan de ses matchs, qui auront permis aux innombrables communautés de Frioulans dispersés dans le monde[7] de renouer avec la mère patrie et la *marilenghe* (la « langue maternelle »). Par ailleurs, échapper aux modèles dominants, y compris aux circuits dominants de la communication, peut être moins un acte politiquement subversif que la manifestation du désir d'une communication plus qualitative, humaine, charnelle. Le formidable succès, il y a quelques années, de la lecture intégrale par relais citoyen de la Bible en langue frioulane est un exemple extraordinaire de cette volonté, de ce besoin viscéral de pratiquer la parole dans un cadre collectif, « national », en présentiel et en usant du corps physique comme résonateur d'identité linguistique et de présence au monde.

Lorsque l'on aborde l'univers des langues minoritaires et ultra-minoritaires on ne peut pas faire l'économie de ce type de besoin humain, profondément humain, de communication incarnée, de proximité physique, qui est d'ailleurs le fondement du critère de réalité dans le discours (Lafont 2007 [1994] : 11-12). Si ce besoin s'exprime dans n'importe quelle langue, il n'en demeure pas moins qu'une langue minoritaire, *a fortiori* une langue ultra-minoritaire, est une langue le plus souvent choisie. Cela veut dire qu'elle ne va pas de soi. Elle n'est pas non plus indispensable : on peut très bien vivre en castillan à Barcelone et même à Gérone, tout en étant enveloppés par le catalan ; on peut vivre en italien à Tolmezzo, à Dronero ou à Maschito, tout en étant côtoyés par le frioulan, l'occitan et l'arbëresh.

[7] On calcule que pour un Frioulan habitant aujourd'hui le Frioul-Vénétie Julienne, il en existe huit ailleurs, notamment dans le continent américain.

Autrement dit, la plupart des usagers des langues minoritaires et tout particulièrement des langues ultra-minoritaires sont des sujets « militants », ou *a minima* volontaires, car, du moins en Occident, ils maîtrisent dans l'écrasante majorité des cas également (et d'abord) la langue d'État, qui est pour eux langue de scolarisation, langue de la vie publique et, souvent, langue première même à la maison.

Cette remarque est fondamentale. Le thème du choix et, plus généralement, le thème du lien à la langue est un dossier riche et passionnant. Nous allons bientôt lui consacrer une monographie restituant les résultats les plus significatifs d'une enquête sur l'élection de l'occitan en France au XX^{e} siècle de la part d'écrivain-e-s et d'activistes. Faute d'espace, nous ne pouvons pas donner d'exemples détaillés ici[8], mais juste rappeler que, pour les auteurs sollicités, les raisons du choix de l'occitan sont très diverses : ici, la pratique courante de l'occitan s'accompagne de la reprise de pratiques agricoles familiales délaissées (Jean-Paul Creissac), alors que là il s'agit d'un témoignage de l'« irréductibilité de l'être humain » (Alem Surre-Garcia) ; ici, la réappropriation de l'occitan vient satisfaire un besoin profond de combler un vide dans la mémoire collective (André Lagarde), ailleurs il s'agit d'une véritable « adoption » ou élection linguistique et culturelle dans une perspective éminemment créatrice (Jean-Claude Forêt)… En bref, chaque auteur a sa propre motivation, précisément parce que, comme souligné plus haut, une langue minoritaire « ne va pas de soi » : son emploi n'est justifié qu'à partir d'un fort ancrage émotionnel, subjectif.

Dans cette perspective, où la subjectivité, l'histoire et la mémoire individuelles ont tant de place ; où la langue minoritaire est pratiquée précisément parce qu'elle se lie intimement avec le sujet ou « être de langage » (Lafont 2004) ; où la langue minoritaire possède une forte portée identitaire en ce qu'elle crée très rapidement de la reconnaissance entre ceux qui la pratiquent et en ce qu'elle se démarque de la langue et des discours dominants… on peut se demander s'il reste une place, et si oui laquelle, pour l'IA, qui finalement ne peut exister qu'à partir de deux conditions nécessaires et insuffisantes :

a) une forte standardisation des langues ;
b) la disponibilité de très grands corpus textuels dans ces langues et dans un nombre suffisamment large de domaines d'usage.

[8] Quelques-uns de ces exemples sont cependant creusés et restitués dans un court article d'il y a quelques ans (Agresti 2011).

Or, ces deux conditions sont précisément ce qui fait le caractère impersonnel d'une langue et le statut d'une langue de grande diffusion ! Comment, dès lors — et, en amont, pourquoi ? — concilier langue (ultra-)minoritaire et IA ? Une fois de plus, le ménage est à tout le moins compliqué… Voyons cela de près.

En linguistique, l'IA est l'horizon contemporain et prochain d'une histoire surtout récente d'outillage technologique des langues naturelles. Nous proposons de distinguer surtout deux champs d'application, qui en réalité se recoupent et entremêlent assez largement :

1) un champ « interne » à la langue et au discours, concernant le traitement numérique (d'abord et surtout de l'écrit : traduction automatique ou assistée par ordinateur ; correcteurs orthographiques et syntaxiques ; dispositifs d'aide à la rédaction sur supports divers ; mais également de l'oral : transcription de l'oral…) ;
2) un champ « externe » à la langue et au discours, concernant le repérage en ligne de ceux-ci, le plus souvent à des fins commerciales ou politiques, dans la Toile (cookies, profilage, etc.) ou dans d'autres contextes (reconnaissance vocale pour activer des dispositifs portatifs ou fixes, des enceintes, etc.).

Cette schématisation rudimentaire permet tout de même de cerner et vérifier *a minima* la nature du rapport liant IA et langues naturelles. La distinction présentée plus haut (Par. 1) s'impose : s'il existe bien une importante *scripta* (grands, voire très grands corpus) en catalan et en basque, s'il existe par ailleurs des institutions de normativisation de ces langues, s'il existe enfin une masse critique d'usagers de ces deux langues « minoritaires », il en va tout autrement pour les langues ultra-minoritaires : peu d'usagers (y compris dans la communication en ligne), forte variation (macro et micro), normes graphiques instables ou polynomiques[9], pas ou fort peu d'usage de la langue dans l'espace public… Par conséquent, on peut douter que la promotion, dans ces contextes mar-

[9] La notion de « langue polynomique » a été utilisée la première fois dans le contexte de l'aménagement de la langue corse mais a été vite appliquée à d'autres contextes linguistiques minoritaires. « [Les langues polynomiques sont des] langues dont l'unité est abstraite et résulte d'un mouvement dialectique et non de la simple ossification d'une norme unique, et dont l'existence est fondée sur la décision massive de ceux qui la parlent de lui donner un nom particulier et de la déclarer autonome des autres langues reconnues. » (Marcellesi 1985 : 314).

qués, de l'IA, soit d'une part faisable et, d'autre part, réellement efficace pour conjurer leur « extinction numérique »[10].

Pourtant, la question reste ouverte, surtout si nous mobilisons la notion, essentiellement dynamique, de « minorisation » (Par. 1.b) au lieu de nous en tenir au statut cristallisé de « (langue) minoritaire » : l'introduction de l'IA dans les contextes minoritaires pourrait en effet permettre de « retrousser la diglossie » (Lafont 1984), c'est-à-dire de modifier en positif les représentations sociales des langues (minoritaires ou minorisées en l'occurrence) pour en favoriser de nouvelles pratiques, davantage « normales »[11]. En effet, toute action, notamment de politique publique, visant à intervenir sur le statut d'une langue (*via* une loi ou bien une action culturelle, y compris la publication d'un dictionnaire ou d'une grammaire de la langue ou la réalisation d'un film en langue locale…) contribue *ipso facto* à modifier, un tant soit peu, les pratiques langagières, ne serait-ce que par amélioration des représentations sociales endogènes et/ou exogènes.

Autrement dit, outiller technologiquement (et par la technologie numérique la plus avancée) une langue peu répandue pourrait rapidement projeter celle-ci dans la contemporanéité et contribuer à la normaliser, les fonctions de la langue minoritaire se rapprochant dès lors de celles de la langue hégémonique. Par ailleurs, l'image de ces langues périphériques serait rehaussée par leur injection dans l'espace public et le maillage social. Bien entendu, ces stratégies ont leur part de volontarisme, qui peut être même fort décalé par rapport à la réalité. Ainsi, l'Agence régionale de la langue frioulane (ARLEF) a financé la publication d'une revue scientifique en version bilingue, anglais-frioulan, et ce sans qu'une claire demande en

[10] Nous tenons cette formulation d'un article paru au Canada dans *Le Devoir* et republié dans le *Courrier international* le 8 novembre 2012. Finalement, l'auteur, Fabien Deglise, se penche sur vingt et une langues européennes, à risque de disparition dans les réseaux numériques « en raison d'une *"prise en charge"* faible ou inexistante de ces langues par les outils de communication qui se déploient actuellement dans le quotidien des gens ». https://www.courrierinternational.com/article/2012/11/08/sauver-les-langues-de-l-extinction-numerique. Dernière consultation : 11 septembre 2021.

[11] C'est le sens qu'une partie croissante de la communauté scientifique francophone accorde à la notion de « normalisation » d'une langue (établir un usage normal, quotidien et aussi public de la langue minoritaire), par rapport à celle de « normativisation » (aménager le corpus d'une langue, en la dotant de « règles » grammaticales, d'un standard orthographique et ainsi de suite).

ce sens ne soit formulée de la part de la communauté linguistique et/ ou scientifique ; aussi, quelques pionniers de la désaliénation culturelle de l'occitan ont-ils pu fonder dans les années 1970 du siècle dernier la revue de sociolinguistique *Lengas*, contenant de manière systématique des contributions scientifiques en langue d'oc. Ces opérations permettent de mettre à jour les langues minoritaires, de faire de la néologie aménagée (Agresti, Puolato 2020) tout en contribuant à en faire des langues en quelque sorte « utiles », dignes de la science, contrairement au poncif incarné dans le célèbre mot de Renan :

> Toute sa vie, on aime à se rappeler la chanson en dialecte populaire dont on s'est amusé dans son enfance ; mais on ne fera jamais de science, de philosophie, d'économie politique en patois.

La discussion reste donc ouverte. Finalement, il appartient aux communautés linguistiques concernées de décider s'il faut emprunter la voie de la pleine standardisation, de la production de *big data*, de la constitution de grands corpus numérisés… préalables incontournables du traitement de la langue par l'IA. Oui, malgré tout, les langues minoritaires peuvent faire bon ménage avec l'intelligence artificielle. Mais c'est une possibilité, pas une fatalité, et encore moins une obligation : c'est pourquoi nous avons formulé quelques réserves à ce propos, en soulignant le particulier statut des langues minoritaires, et notamment celles très peu répandues, qui probablement résistent précisément parce qu'elles sont foncièrement différentes des autres langues naturelles et répondent à d'autres besoins du sujet et du groupe. Après tout, avant même que de parler de (re)fonctionnalisation des langues minoritaires, le plus important, c'est que celles-ci possèdent ne serait-ce qu'une fonction (identitaire, cryptique, affective, etc.) pour qu'on leur accorde une raison d'être et de perdurer. Pour que quelqu'un continue, malgré tout, de les choisir et pratiquer.

Conclusion et ouverture de perspective

En résumant, la nature du rapport entre l'IA et les langues minoritaires dépend en bonne mesure du réel statut sociolinguistique et démolinguistique de celles-ci. Pour en rester au contexte européen, une chose est le tchèque ou le néerlandais, autre chose est le catalan ou l'irlandais, autre chose encore est le walser ou le yiddish : des variables comme la taille des corpus linguistiques numérisés disponibles, la nature et la dimension de

la *scripta*, la variation micro et macro, le niveau de normativisation et standardisation et la volonté des communautés concernées déterminent la faisabilité du développement de l'IA dans ces contextes linguistiques divers. Le cadre de communication et les enjeux que celui-ci sous-tend jouent également un rôle incontournable dans cette évaluation : le multilinguisme au sein des institutions européennes est une configuration qui très souvent reste uniquement sur le papier, les textes officiels, les appels d'offres, les comptes rendus des travaux des différentes commissions, etc. étant le plus souvent publiés juste en une, deux, trois langues (anglais et/ ou français et/ ou allemand)[12]. Dans ces conditions, des langues officielles peuvent devenir, *de facto*, des langues minoritaires. Par ailleurs, le statut de langue (ultra-)minoritaire peut évoluer en raison d'aménagements linguistiques positifs ou négatifs, et la minorisation peut être compensée ou endiguée par un processus de normalisation. Aujourd'hui, face aux hégémonies en place, et notamment face au rouleau compresseur de l'anglais *lingua franca* internationale, on a raison de se poser la question suivante : « nos langues vont-elles toutes devenir des *heritage languages* ? » (Anquetil, Vecchi 2018). On ne peut guère parler de l'IA dans le cadre et en fonction de l'« intégration européenne »[13] en faisant fi de ces questionnements, qui montrent bien la primauté des politiques linguistiques sur l'outillage technologique des langues européennes.

Ce dernier scénario (une langue de service internationale hégémonique au point de réduire toutes les autres langues, y compris nationales, au statut de langues patrimoniales ou d'héritage) pose un gros problème éthique qui met en danger la possibilité même d'une pleine intégration européenne. Il est loin d'être invraisemblable et on aurait tort de l'ignorer ou de le sous-estimer. Dans cette perspective, le plus important est d'éviter des dérives « religieuses » ou déterministes : la primauté d'une langue naturelle n'est en aucun cas à attribuer à des forces surnaturelles, à des hypostases telles que l'Histoire, la Science ou l'Homme. De même que les langues ne sont pas intrinsèquement porteuses de valeurs spécifiques (Bourdieu 1982). Non, ce n'est pas l'Histoire qui a décidé de l'hégémonie anglophone, dans la communication scientifique et au-delà, comme le prétendent d'aucuns en usant d'un raccourci douteux : c'est plutôt la politique qui a pris des

[12] En effet, même lorsque l'on dispose de plusieurs versions d'un même texte, il n'y en a que très peu qui possèdent le statut de texte officiel, faisant foi, et non de traduction.

[13] Nous convoquons là l'intitulé du projet de recherche qui inspire le présent débat.

décisions souvent irrationnelles, contre-productives pour les économies nationales autres que britanniques ou états-uniennes et liées à des moments de l'histoire très précis, tout particulièrement à partir de la fin de la Seconde Guerre mondiale (Tremblay 2018). C'est le retour du réflexe de Babel, pour lequel la diversité linguistique est moins une richesse qu'un problème (une « barrière »), réflexe décliné aujourd'hui à l'aune de la globalisation marchande, dont le paradigme finit par façonner ou du moins lourdement conditionner nos vies au quotidien. De même, ce ne sont pas les citoyennes et citoyens d'Europe qui ont choisi l'anglais, comme le proposait Tullio De Mauro (2014), et ce pour au moins deux bonnes raisons : a) il n'y a jamais eu de référendum européen en ce sens ; b) dans le scénario d'une Europe tendanciellement monolingue anglophone le TAL (« taux d'aliénation linguistique ») serait particulièrement élevé (Gazzola 2016). Comment bâtir une solide « démocratie européenne » sur la mise à distance des majorités populaires par rapport aux institutions continentales, déjà plutôt mal perçues par la doxa et fragilisées par des mouvements politiques eurosceptiques voire carrément anti-européens ?

Le régime multilingue qui caractérise l'Union Européenne et qui fait de celle-ci un formidable espace pluraliste doit être défendu précisément dans une perspective de démocratie, aussi participative et inclusive que possible. C'est une question de droits linguistiques, donc de droits humains (Agresti 2021b). Certes, il faut outiller ce multilinguisme pour qu'il soit effectif et efficace. Mais la bonne nouvelle est que la pleine démocratie linguistique en Europe coûte beaucoup moins cher que ne le prétendent les discours superficiels ou les croyances diffuses : un peu plus de deux euros l'an *pro capite* (Gazzola 2016 : 284) ! Alors qu'un régime mono ou oligolingue serait infiniment plus coûteux (et injuste) pour les citoyens du Vieux Continent[14]. Et cela pourrait coûter encore moins cher grâce aux

[14] Imaginons un appel d'offres européen, publié uniquement en langue anglaise. Une association de Roumanie, mettons, pourrait ne pas avoir en son sein de personnel en mesure de le comprendre à la perfection et, *a fortiori*, elle pourrait ne pas disposer de ressources humaines en mesure de rédiger dignement le projet en anglais pour répondre à cet appel. Cette association devrait donc passer par les services d'un traducteur, ce qui risque de lui coûter très cher et ce qui signifie disposer de beaucoup moins de temps pour la rédaction du projet, car celui-ci ne pourra être validé qu'une fois traduit. Enfin, malgré l'argent dépensé, cette association n'aurait aucune garantie de la qualité de la traduction : un traducteur professionnel, s'il est censé connaître les langues avec lesquelles il travaille, peut ignorer tout du fond du projet et donc, fatalement, très mal le traduire. On voit bien que l'adoption d'un régime mono- ou oligolingue en Europe entraîne une avalanche d'effets néfastes. Et les problèmes ne s'arrêtent pas là… (Frath 2018).

progrès de la traduction automatique, donc grâce au potentiel de l'IA appliqué au « champ interne » à la langue et au discours (voir *supra*, Par. 3.1). Voilà un aspect positif du ménage entre l'IA et les langues (relativement) minoritaires de l'UE. En raison de ce particulier développement technologique les vingt-quatre langues officielles[15] pourraient garder leur statut *de jure* et *de facto* et on pourrait même en introduire d'autres dans des contextes spécifiques. Mais, nous insistons sur ce point, c'est toujours la politique qui décide.

Ces considérations suggèrent que la vraie « guerre des langues » n'est pas entre les langues, bien évidemment. La formule à succès de Calvet, c'est de la poudre aux yeux. On devrait parler plutôt de « guerre (politique, économique, symbolique…) par les langues ». Et par les discours, bien entendu. Onésime Reclus au début du XX^{e} siècle pour la francophonie (Agresti 2021a) et Winston Churchill dès les années 1940 pour l'anglophonie avaient très bien compris que la domination linguistique et culturelle rapporterait beaucoup plus de bénéfices et entraînerait beaucoup moins de risques que les conquêtes militaires. Ainsi, la conquête linguistique, déguisée sous mille beaux masques (prestige littéraire, exportation de la démocratie et de styles de vie séduisants, libéralisation du commerce, essor des industries créatives — chanson et cinéma d'abord et surtout —, etc.) intègre d'autres conquêtes : culturelle au sens large, économique, politique. Qui plus est, ces conquêtes se font souvent avec le consentement des populations conquises, bientôt acquises à la culture dominante.

Or, il y a lieu de croire que l'IA participe de ce processus de conquête, de colonisation des esprits, mais d'une manière très subtile, voire sournoise. En effet, si au niveau du « champ externe » (voir *supra*, Par. 3.2), elle permet de contrôler, y compris de manière policière, ce qui se dit dans les réseaux numériques en ligne et de profiler la communication à destination des clients/ électeurs, donc de « fabriquer le consensus » (Chomsky, Herman 1988), au niveau du « champ interne » la traduction automatique reconduit et met à jour ce caractère « fasciste » que Roland Barthes prêtait, non sans provocation, aux langues naturelles (Barthes 1978). De quoi s'agit-il exactement ? Une anecdote amusante peut être utilement convoquée à ce sujet.

[15] Nous rappelons que, même depuis que le Brexit est devenu une réalité, l'anglais « constitue toujours une langue officielle de l'Union européenne et une langue de travail, tant que le Conseil de l'Union européenne ne se prononce pas, à l'unanimité, pour la retirer » (https://www.touteleurope.eu/fonctionnement-de-l-ue/les-langues-de-l-union-europeenne-en-3-minutes/. Dernière consultation : 16 septembre 2021).

La première mouture du présent texte avait été rédigée oralement, pour des raisons purement circonstancielles et pratiques mais aussi pour tester cette particulière fonctionnalité de l'IA, la reconnaissance vocale (RV) associant capacité de reconnaître le langage naturel à l'oral et référence à des corpus textuels de grosse taille. Or, nous avons arrêté cette première rédaction lorsque, peu après avoir commencé à dicter notre propos (et ce de la manière la plus claire, standardisée et intelligible), le système de RV de Google a transcrit trois fois « en allemand » au lieu de « en amont », trois fois répété…

La langue est « fasciste » (l'allemand n'y est pour rien !) et l'IA ajoute une nuance à cette provocation. Sans vraiment l'écouter, elle force l'auteur/le locuteur/ l'écrivain à se ranger du côté des majorités et/ou des discours majoritaires ou plus exactement plus probables, prévisibles, dans les langues de plus large diffusion. Elle induit en quelque sorte un conformisme discursif, statistiquement hégémonique, qui est d'ailleurs sa principale référence. La question se pose alors de savoir si, à terme, la diffusion de l'IA ne finirait pas par (contribuer à) brider notre créativité et notre originalité de pensée, un peu comme il arrive à des écrivains d'être réécrits par des experts des maisons d'édition afin de conformer leurs œuvres aux goûts présumés du public…[16]

Le conformisme marchand qui en arrive à homologuer, encore que partiellement, l'expression par excellence de la subjectivité créatrice est un exemple qui mérite toute notre attention. On peut renverser la donne : la littérature est précisément convoquée par Umberto Eco au moment de questionner la provocation barthésienne. Chez le sémiologue italien le jeu littéraire est une véritable « tricherie » en mesure de casser la circularité emprisonnante du discours, sa prévisibilité et sa banalisation (Eco 1979). En effet, on a pu parler des chefs-d'œuvre littéraires comme de véritables erreurs éditoriales.

Or, cela risque d'être d'autant plus vrai lorsque l'outil même de la création littéraire est un outil rare, original en soi, peu courant et accessible : voici que la littérature en langue minoritaire, et *a fortiori* ultra-minoritaire, pourrait représenter un vrai et surprenant espace de liberté. C'est

[16] À ce sujet, la position critique de Bernard Stiegler, formulée à différentes occasions, à l'égard des *big data* et de leur exploitation peu éclairée, nous paraît tout à fait salutaire. L'IA doit être soumise à la raison et à la créativité humaines et non l'inverse, autrement nous nous heurterons à une reconduction infinie de l'état de fait.

que la condition de langue (ultra-)minoritaire condamne le plus souvent les textes qui l'actualisent à la marginalisation dans le cadre du marché. L'enjeu de la création littéraire se situe ailleurs. Pour paradoxal que cela puisse paraître, le statut de la création en langue (ultra-)minoritaire finit par se rapprocher de celui de la création d'avant-garde, libérée du souci du consensus et de la rentabilité économique. Comme le rappelait le provençal Robert Lafont (1923-2009) lors d'un entretien d'il y a une quinzaine d'années, au sujet du choix presque exclusif de l'occitan par rapport au français dans ses œuvres littéraires[17], « lorsqu'un écrivain a un public de cent cinquante lecteurs très fidèles, il est heureux »[18]. Par ailleurs, on put reprocher à l'écrivain gascon Bernard Manciet (1923-2005) d'avoir intentionnellement « inventé » sa langue, le parler noir des Landes de Gascogne déjà si marqué et même aberrant par rapport aux formes standardisées de l'occitan du sud-ouest. Mais, on le sait bien, les auteurs finissent par faire autorité. Preuve en est, cette fois-ci en contexte languedocien, Jean Boudou (1920-1975), qui incarne le paradoxe d'un écrivain à la plume et à la vie très singulières, même mystérieuses, devenu pourtant l'un des classiques de la littérature d'oc contemporaine.

On se soucie de « sauver » les langues menacées de disparition, à l'instar d'espèces animales ou végétales, sans trop articuler ce programme. Pourtant, ce sont ces langues qui, en quelque sorte, nous sauvent, en nous permettant d'être et de nous dire au monde de la manière la plus adaptée à nos nécessités. Ce qu'on doit protéger, c'est précisément la diversité linguistique moins en tant que valeur en soi, abstraite, qu'en tant que richesse de possibilités et ressources expressives, relationnelles, cognitives. Il faut protéger et assurer le droit de se servir, en privé comme en public, de notre langue maternelle ou d'élection. Il faut par ailleurs protéger le droit et le devoir d'accéder à la mémoire collective, en assurant la perdurance des conditions de transmission/réception des langues qui véhiculent cette mémoire (orale ou écrite). Si nous sommes capables de faire cela, c'est notre liberté d'être pensants, de réservoirs de mémoires et d'émotions que nous sauvegarderons et valoriserons.

[17] Rappelons au passage que Robert Lafont a presque uniquement utilisé le français dans ses ouvrages scientifiques et presque exclusivement l'occitan dans ses œuvres littéraires.

[18] Cette citation est tirée du film *Robert Lafont. Un écrivain dans le siècle*, de Christian Passuelo (2001). http://www.film-documentaire.fr/4DACTION/w_fiche_film/8950_1. Dernière consultation : 11 septembre 2021.

***

Post-scriptum. La liberté offerte par la marge, la *talvèra* de l'épigraphe[19], signifie aussi la possibilité d'échapper au contrôle et à la manipulation du discours rendu possible par la centralisation de la communication en ligne : une infox passe difficilement en occitan ; une arnaque en ligne ne sera sans doute jamais montée en usant du croato-molisain ou du cornique. L'usage de langues minoritaires peut en ce sens limiter les mille violences et supercheries dont l'univers du numérique — redevenu village aussi en raison de la centralisation outrée réalisée par l'IA — est semé. Dès lors, la diversité linguistique prend une tout autre connotation : de barrière linguistique à barrage. La différence est capitale.

[19] *Talvèra* indique en occitan la marge du champ qui ne peut pas être sillonnée et donc ensemencée. C'est une lisière où « sont tolérés chiendent et herbes dites mauvaises ». https://latalvere.org/cest-quoi-ce-mot/. Dernière consultation : 11 septembre 2021.

Bibliographie

Agresti Giovanni (2021a). « Francophonie et multilinguisme en contexte africain : du conflit culturel à la coopération ». *Mondes et cultures*, revue de l'Académie des Sciences d'Outre-Mer, 1, 467-486.

Agresti, Giovanni (2021b). « Droits linguistiques ». In : Josiane Boutet, James Costa (Sous la direction de), *Dictionnaire de la sociolinguistique, Langage et société,* 2021/HS1, Editions de la Maison des sciences de l'homme, p. 115-118. En ligne : https://www.cairn.info/revue-langage-et-societe-2021-HS1-page-115.htm

Agresti Giovanni (2017). « Le web et les langues minoritaires. L'exemple du corse et de l'occitan ». In : Giovanni Agresti. *Du centre et de la périphérie. Au carrefour d'italophonie et francophonie*. Préface d'Henri Giordan. Rome : Aracne (« L'essere di linguaggio », 4), 215-233.

Agresti Giovanni (2016). « Nous sommes tous minoritaires ! Besoins de médiation et malaise linguistique ». *ELA. Études de linguistique appliquée*. 181, numéro thématique « Médiation et droits linguistiques » (sous la direction de Giovanni Agresti, Michele De Gioia avec la collaboration de Mario Marcon). Paris : Didier Érudition/ Klincksieck, 79-92.

Agresti Giovanni (2011). « Le sujet reconfiguré, le marché linguistique redessiné : douze ans d'enquête sur le lien à la langue (1996-2008) ». In : Angelica Rieger, Domergue Sumien (éds). *L'Occitanie invitée de l'Euregio. Liège 1981 – Aix-la-Chapelle 2008 : Bilan et perspectives*. Actes du Neuvième Congrès International de l'Association Internationale d'Études Occitanes, Aix-la-Chapelle, 24-31 août 2008. Aachen : Shaker, 597-603.

Agresti Giovanni, Puolato Daniela (2020). « Mettre à jour une langue minoritaire. Le francoprovençal des Pouilles : stratégies et enjeux néologiques ». *Neologica*, 14, 61-82.

Anquetil Mathilde, Vecchi Silvia (2018). « Nos langues vont-elles toutes devenir des *Heritage Languages* ? ». In : Giovanni Agresti, Joseph-G. Turi (éds). *Du principe au terrain. Norme juridique, linguistique et praxis politique*, Actes du Premier Congrès mondial des droits linguistiques, Vol. II. Rome : Aracne (« Lingue d'Europa e del Mediterraneo / Diritti linguistici », 14), 277-295.

Barthes Roland (1978). *Leçon*. Paris : Seuil.

Bourdieu Pierre (1982). *Ce que parler veut dire. L'économie des échanges linguistiques*. Paris : Fayard.

Calvet Louis-Jean (1999a). *La guerre des langues et les politiques linguistiques*. Paris : Hachette.

Calvet Louis-Jean (1999b). *Pour une écologie des langues du monde*. Paris : Plon.

Chomsky Noam, Herman Edward S. (1988). *Manufacturing Consent : The Political Economy of the Mass Media*. New York : Pantheon Books.

De Mauro Tullio (2014). *In Europa son già 103. Troppe lingue per una democrazia?* Bari : Laterza.

Eco Umberto (1979). « La Lingua, Il Potere, La Forza ». *Alfabeta*, I, 1 (mai), 2-3.

Foucault Michel (1975). *Surveiller et punir. Naissance de la prison*. Paris : Gallimard (« Bibliothèque des histoires »).

Frath Pierre (2018). « Une classification gnoséologique des langues au service de la politique linguistique ». *Repères-DoRiF*, 17, numéro thématique « Diversité linguistique, progrès scientifique et développement durable ». https://www.dorif.it/reperes/pierre-frath-une-classification-gnoseologique-des-langues-au-service-de-la-politique-linguistique/ (Dernière consultation : 11 septembre 2021).

Gazzola Michele (2016). « Multilinguisme et équité: l'impact d'un changement de régime linguistique européen en Espagne, France et Italie ». In : Giovanni Agresti, Joseph-G. Turi (sous la direction de). *Représentations sociales des langues et politiques linguistiques. Déterminismes, implications, regards croisés*, Actes du Premier Congrès mondial des droits linguistiques, Vol. I. Rome : Aracne (« Lingue d'Europa e del Mediterraneo/ Diritti linguistici », 12), 269-286.

Julia Luc (2020). *L'intelligence artificielle n'existe pas*. Paris : J'ai lu.

Lafont Robert (2007 [1994]). *Il y a quelqu'un. La parole et le corps*. Limoges : Lambert-Lucas.

Lafont Robert (2004). *L'être de langage. Pour une anthropologie linguistique*. Limoges : Lambert-Lucas.

Lafont Robert (1984). « Pour retrousser la diglossie ». *Lengas, revue de sociolinguistique*, 15, 5-36.

Marcellesi Jean-Baptiste (1985). « La définition des langues en domaine roman : les enseignements à tirer de la situation corse », In : AA.VV. *Actes du Congrès de Linguistique Romane de 1983*, vol. V, « Sociolinguistique », Université d'Aix-en-Provence, 307-314.

Morin Edgar (1986). *La méthode*, tome 3. *La Connaissance de la Connaissance*. Paris : Seuil.

Poggeschi Giovanni (éd) (2012). *Le iperminoranze*. San Cesario di Lecce : Pensa editore.

Quenot Sébastien (2010). « Le corse dans la cyberguerre mondiale des langues ». In : Giovanni Agresti, Mariapia D'Angelo (éds). *Renverser Babel. Économie et écologie des langues régionales et minoritaires*. Actes des Troisièmes Journées des Droits Linguistiques (Teramo-Faeto, 20-23 mai 2009). Rome : Aracne, 345-362.

Shah Idries (2009). *Apprendre à apprendre*. Paris : Le Courrier du Livre (« Collection soufisme vivant »).

Tremblay Christian (2018). « L'enseignement supérieur et la recherche entre anglicisation et internationalisation ». *Repères-DoRiF*, 17, numéro thématique « Diversité linguistique, progrès scientifique et développement durable ». https://www.dorif.it/reperes/christian-tremblay-lenseignement-superieur-et-la-recherche-entre-anglicisation-et-internationalisation/ (Dernière consultation : 11 septembre 2021).

De Europa Première partie - Special Issue 2022

Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive

Guido Vetere

Introduzione

Le tecnologie per il trattamento del linguaggio naturale sono un ingrediente fondamentale di quella trasformazione detta "digitalizzazione" che oggi coinvolge le società su scala globale. Nell'ultimo decennio, queste tecnologie, a cui ci si riferisce in generale con la sigla HLT (*Human Language Technologies*) o, più specificamente, NLP (*Natural Language Processing* o Elaborazione del linguaggio naturale) hanno fatto notevoli progressi grazie soprattutto alle crescenti capacità dell'intelligenza artificiale (IA) applicate al materiale linguistico. Nel 2020 le tecnologie linguistiche hanno compiuto un significativo balzo in avanti per merito di OpenAI, una società di ricerca non-profit della Silicon Valley fondata nel 2015 da Elon Musk e oggi finanziata tra gli altri da Microsoft, la cui missione dichiarata è quella di "democratizzare" l'intelligenza artificiale (*Artificial Intelligence*, AI), cioè di renderne i benefici accessibili a tutti. Nel maggio del 2020, in piena crisi pandemica, OpenAI ha rilasciato la terza versione del suo *Generative Pre-trained Transformer* (GPT-3), un sistema neurale capace di rispondere a domande, sviluppare temi, dialogare e tradurre, in modo spesso indistinguibile dagli esseri umani (Brown *et alii* 2020). Non solo le prestazioni del sistema sono sorprendenti, ma sono ottenute senza doverlo addestrare per ciascuno specifico scopo: si tratta quindi, secondo alcuni commentatori, di un'intelligenza linguistica generale più simile a quella umana rispetto alle precedenti. L'addestramento di GPT-3 ha richiesto risorse computazionali ed energetiche del costo di milioni di dollari, ma le prestazioni del sistema sono tali che gli stessi sviluppatori hanno messo in guardia verso i possibili usi illeciti che se ne potrebbe fare. In effetti, poche settimane dopo il lancio, uno studente californiano aveva già aper-

Guido Vetere, Università degli Studi Guglielmo Marconi, g.vetere@unimarconi.it

to un blog generato automaticamente da GPT-3. Non molto più tardi, il *Guardian* pubblicava, come monito, un finto articolo scritto da quella AI[1].

Le tecnologie del linguaggio naturale rappresentano uno dei principali settori dell'IA, non solo per la loro immensa potenzialità economica, ma anche per il fatto di affrontare direttamente la materia dell'intelletto umano. Tutte le tecnologie intelligenti hanno qualcosa a che fare con la cognitività della nostra specie: si pensi ad esempio alla guida autonoma, dove è importante riuscire a classificare le forme visibili dal veicolo in modo simile a quanto fa normalmente chi è al volante. Ma il linguaggio è il modo stesso in cui rappresentiamo nella coscienza gli oggetti che i sensi ci consegnano, ed anzi, secondo alcune ipotesi, entra direttamente nel processo di identificazione di tali oggetti. Non è esagerato dire che il programma dell'intelligenza artificiale, delineato da Alan Turing negli anni '50 del XX secolo, non potrà essere realizzato senza la creazione di una piena intelligenza linguistica.

L'impiego delle moderne tecniche di NLP include la ricerca e la classificazione di informazione per la gestione dei contenuti prodotti sul web o all'interno delle imprese, e in generale per la *business intelligence,* con significative applicazioni nella sanità; la traduzione e correzione automatica per il commercio elettronico, l'editoria, le relazioni istituzionali; l'analisi e classificazione di testi anche frammentari per la sicurezza, i media sociali e la profilazione dell'utenza; i servizi di interrogazione e dialogo per il supporto alla clientela, le relazioni col pubblico e l'assistenza personale; la generazione o la trasformazione di testi per gioco e intrattenimento, per editoria, o il lavoro creativo; la robotica e l'automazione, anche nel settore automobilistico.

Il mercato delle tecnologie di NLP è dunque in forte espansione: gli analisti di *Fortune* prevedono, entro il 2026, una crescita dell'indice CAGR (*Compound annual growth rate*) del 29,4%[2], che colloca il settore tra quelli maggiormente attrattivi per gli investimenti. L'impatto delle tecnologie del linguaggio naturale diviene sempre più significativo nella vita sociale,

[1] "*A robot wrote this entire article. Are you scared yet, human?*", *The Guardian*, 8 settembre 2020, https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3 (data ultima consultazione: 29/09/2021).

[2] Rapporto *Fortune*, giugno 2021: "*The global natural language processing market is projected to grow from $20.98 billion in 2021 to $127.26 billion in 2028 at a CAGR of 29.4%*", https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933 (data ultima consultazione: 29/09/2021).

non solo per le piattaforme che le usano intensamente (talvolta con finalità opache), ma anche per le applicazioni a scopi di utilità pubblica come, ad esempio, quelle recentemente sviluppate per il contrasto alla pandemia da COVID-19[3].

Per la crescente pervasività ed il ruolo sempre più incisivo che le tecnologie del linguaggio giocano nei sistemi economici e politici, il tema del loro sviluppo nelle diverse aree linguistiche assume un carattere di rilevanza strategica. Il disequilibrio di queste tecnologie, ed in particolare la preminenza, in esse, della lingua inglese, entra infatti a far parte del problema geopolitico riguardante i monopoli tecnologici: le concentrazioni di dati e capacità computazionali hanno raggiunto dimensioni tali da produrre distorsioni negli equilibri economici e sociali. Il possesso di migliori tecnologie linguistiche, in particolare, si traduce in migliori servizi per le aziende e i cittadini, maggiore efficienza delle amministrazioni e delle filiere produttive, più ampie capacità della ricerca pubblica e privata. Un capitolo specifico riguarda l'efficienza dei sistemi di traduzione automatica, dove la differenza qualitativa che oggi ancora si riscontra nelle traduzioni da e per l'inglese è molto significativa. L'Unione europea sta intensificando gli sforzi per colmare questo divario, anche con iniziative di supporto e incentivo alla produzione di risorse linguistiche computabili. La disponibilità di tali risorse, pur importante, non esaurisce tuttavia il problema del divario. Vi sono infatti aspetti legati alla teorizzazione, alle tecnologie di base, alle architetture, alle competenze, ai modelli di sviluppo, che vanno presi adeguatamente in considerazione nel disegno delle politiche di ricerca e sviluppo industriale del settore, sia a livello comunitario sia al livello dei singoli Stati.

Il presente contributo vuole offrire una riflessione sullo stato e le prospettive di ricerca delle tecnologie per lingue diverse dall'inglese, e in particolare per le maggiori lingue dell'Unione europea, con lo scopo di fornire elementi utili per comprendere il quadro attuale e i suoi possibili sviluppi. Forniremo dapprima una breve introduzione alle tecniche che nel corso dell'ultimo decennio hanno rivoluzionato il campo delle tecnologie linguistiche. Questa sarà utile a comprendere il funzionamento della "fabbrica del linguaggio" basata sull'odierna intelligenza artificiale ed il ruolo che le risorse linguistiche giocano al suo interno. Infine, si ragionerà sulle alter-

[3] Allen Institute of AI: COVID-19 Open Research Dataset, https://allenai.org/data/cord-19 (data ultima consultazione: 29/09/2021).

native che, sul piano metodologico e tecnologico, possono contribuire ad uno sviluppo più equilibrato e governabile delle tecnologie linguistiche nel prossimo futuro.

1. La linguistica delle reti neurali

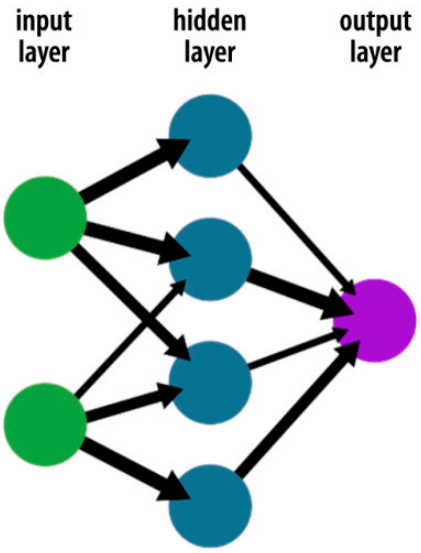


Figura 1. Una semplice rete neurale

Gli sviluppi recenti delle tecnologie linguistiche sono caratterizzati dall'uso pervasivo delle "reti neurali". Queste sono sistemi *software* basati su strutture di elementi atomici (detti appunto "neuroni") disposti su strati (*layer*) funzionalmente distinti in uno strato esterno (*input*), alcuni strati interni (*hidden*) e uno esterno (*output*), dove i neuroni di uno strato sono connessi, con varia forza, a quelli dello strato successivo (Fig. 1). Mediante un processo detto "apprendimento", le connessioni tra neuroni (dette anche "parametri") vengono rafforzate o indebolite fintanto che i dati presentati in *input* non producono, tramite un processo di propagazione, una soddisfacente approssimazione dei dati desiderati in *output*. In sostanza, le reti neurali "imparano" a produrre un *output* a fronte di un *input* mediante la regolazione delle proprie connessioni interne. Questo richiede una specifica codifica dell'*input*, una appropriata architettura interna, e una funzione di calcolo dell'errore nella predizione (*loss function*) la quale, in cicli di addestramento detti "epoche", consente di valutare volta per volta, il progresso del sistema (Goodfellow *et alii* 2016).

Le reti neurali, una volta addestrate e messe in esercizio, funzionano essenzialmente come classificatori: semplificando, si può dire che ciascuna configurazione di dati presentati in *input* produce un simbolo di *output*. Tali sono ad esempio i sistemi di riconoscimento delle immagini oggi presenti in molti *smartphone*, in grado di identificare i volti umani in una fotografia: ciascuna regione dell'immagine è classificata come contenente, o non contenente, un volto. In genere, i dati con cui si opera l'addestramen-

to devono essere pre-elaborati (apprendimento supervisionato), ma in certi casi questo non è necessario (apprendimento non supervisionato). La pre-elaborazione (annotazione) dei dati di addestramento rappresenta uno dei maggiori "colli di bottiglia" delle tecnologie neurali: questa infatti richiede in genere molto lavoro umano, e in alcuni casi assicurarne la qualità è problematico. Per converso, la possibilità di impiegare dati non supervisionati rappresenta una notevole opportunità, posto che siano stati identificati metodi idonei per estrarli da ciò che è disponibile. Questa opportunità si presenta in effetti nelle applicazioni linguistiche grazie al carattere lineare del linguaggio e alla possibilità di applicarvi ragionamenti combinatori. Ad esempio, presa una sequenza di *token* (grosso modo, occorrenze di parole), si può istruire una rete neurale a prevedere (o generare) il *token* occorrente in una certa posizione sulla base dei *token* precedenti e successivi.

Le moderne tecniche di NLP si basano su rappresentazioni puramente numeriche delle unità linguistiche. Ogni unità (ad esempio radici lessicali o morfologiche) è codificata come un vettore di numeri reali (nell'ordine delle centinaia) detti "*embedding*", che le rende confrontabili le une alle altre. L'insieme di *embedding* di una lingua (o una sua proiezione settoriale) è detta "modello linguistico" (*language model*). Una tecnica classica per ottenere tali *embedding* in modo non-supervisionato è detta *skip-gram*[4]: in una sequenza di *token* (parole o frammenti) ne viene rimossa ("mascherata") qualcuna. In fase di addestramento, la sequenza così ottenuta viene data in *input* ad una rete il cui scopo è indovinare, in base al contesto, le parole rimosse. Quando la rete avrà imparato a svolgere il compito, dai suoi parametri, così come saranno stati stimati, verrà estratto, con opportune tecniche, il vettore che ne rappresenta l'*embedding*. In sostanza, la rete impara a riconoscere la correlazione tra le parole in base alla loro distribuzione; da essa, per ciascuna parola, viene poi estrapolata una rappresentazione numerica che ne riflette le proprietà. Molta della attuale ricerca in NLP è finalizzata a trovare il modo ottimale per ottenere questo genere di rappresentazioni (Ferrone, Zanzotto 2020).

L'approccio neurale non supervisionato alla modellazione linguistica viene spesso ricondotto all'"ipotesi distribuzionale" di Zelig Harris (Harris 1954). Tuttavia, le ragioni per la diffusione di questo approccio non

[4] Questa tecnica fa parte di una classe di metodologie analoghe, per una introduzione si rimanda a Manning e Schütze (1999).

vanno cercate nella linguistica del Novecento, ma nell'ingegneria del Duemila. I metodi distribuzionali consentono infatti di ottenere in modo automatico, sulla base dell'enorme quantità di testi oggi disponibili in forma digitale, una rappresentazione delle unità della lingua che si dimostra efficace anche in molti compiti che richiedono valutazioni di tipo semantico, senza peraltro richiedere alcun modello esplicito del significato. Accade in effetti che la distribuzione delle parole nei contesti in cui appaiono faccia emergere fenomeni legati alla sfera concettuale, e che di conseguenza, proiettando gli *embedding* in uno spazio geometrico multidimensionale, le misure di distanza in questo spazio possano rendere conto in qualche modo anche di relazioni che si collocano sul piano semantico (Fig. 2)[5]. Tale circostanza, invero non sorprendente, entusiasma gli ingegneri.

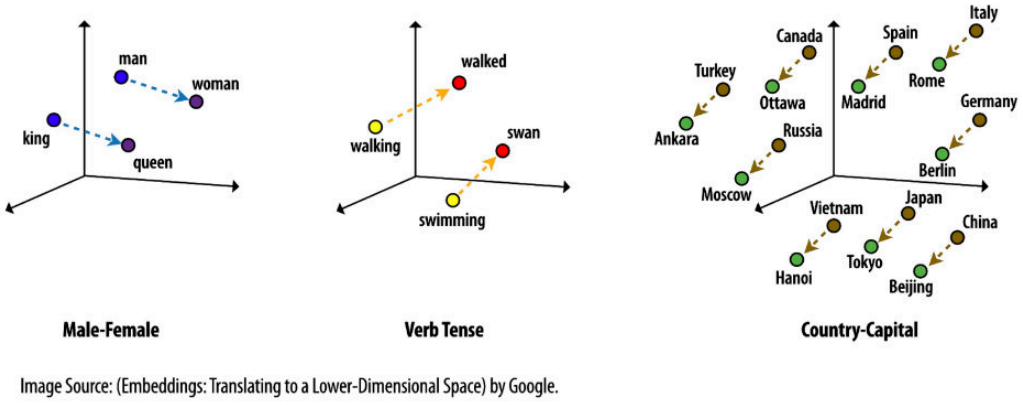


Figura 2: Similarità semantica come distanza nello spazio degli *embedding*

Al di là di alcuni accenni, ad esempio alla filosofia di Wittgenstein[6], l'ingegneria del linguaggio non ambisce in genere ad alcuna teoria del significato, essendo mossa da finalità pratiche e industriali. Ma di fronte a certi successi delle reti neurali, questa "modestia" si capovolge talvolta nell'idea che, nel modo in cui il linguaggio è dato agli automi, qualsiasi ipotesi sul "piano del contenuto" sia dispensabile, e che il distillato statistico della testualità digitale possa rendere superflua la linguistica, *in primis* la lessi-

[5] Una sintetica descrizione del procedimento è contenuta nel corso rapido di Google: *Embeddings: Translating to a Lower-Dimensional Space,* https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space (data ultima consultazione: 29/09/2021).

[6] Intervista ad Amid Singhal, al tempo Vicepresidente di Google, Wired 2010, https://www.wired.com/2010/02/ff-google-algorithm/ (data ultima consultazione: 29/09/2021).

cografia. Se si tralasciano alcune posizioni ingenue sotto il profilo epistemologico[7], non si intravede l'intenzione di fondare una visione "monoplanare" del linguaggio in cui il segno, elemento espressivo che rimanda ad un contenuto concettuale, possa essere ridotto alla combinatoria dei significanti ed eliminato. Tuttavia, nelle tecnologie linguistiche neurali, questa visione prende in effetti corpo. Dunque si può osservare come uno dei principali sviluppi tecnologici che sono oggi alla base delle trasformazioni della cultura materiale non si accompagni ad una adeguata critica dei suoi fondamenti teorici. Questo pragmatismo concorre al vantaggio di chi possiede i mezzi per conseguire le finalità applicative, poiché favorisce le tecnologie di lingue, come l'inglese, dotate di maggiori risorse. Non solo l'inglese si configura come "*lingua franca*" della globalizzazione, ma la sua semplice morfologia la rende ottimale per il trattamento con metodi distribuzionali. Anzi si può ipotizzare che questi ultimi si siano perfezionati proprio in base alle caratteristiche morfosintattiche dell'inglese. Così come, secondo alcuni, la dottrina di Aristotele aveva promosso le forme del greco antico al livello della metafisica[8], le forme dell'inglese sembrano oggi destinate ad impiantarsi nelle piattaforme tecnologiche nelle quali risiedono sempre maggiori porzioni della nostra vita quotidiana.

2. La fabbrica del linguaggio

Il processo di fabbricazione delle tecnologie linguistiche prevede, nella generalità dei casi, un ciclo che parte dalla costruzione, o il reperimento, di modelli linguistici generali, la loro specializzazione rispetto al dominio applicativo, l'integrazione con i sistemi informativi nei quali si collocano (Fig. 3). Oggi sono a disposizione numerose piattaforme *software* aperte in grado di gestire con efficienza processi di questo tipo[9]. Tuttavia, la possibilità per gli automi linguistici di svolgere *task* complessi quali la traduzione automatica, la ricerca di informazione su base semantica, il *question*

[7] Un Manifesto di questo "positivismo dei dati" è l'editoriale di Chris Anderson, "*The end of theory: The data deluge makes the scientific method obsolete*", *Wired*, maggio 2008. https://www.wired.com/2008/06/pb-theory/ (data ultima consultazione: 29/09/2021).

[8] Era questa l'opinione del filologo tedesco Hermann Steinthal (Gröbzig, 16 maggio 1823 – Berlino, 14 marzo 1899).

[9] La lista di piattaforme *open source* che implementano funzioni di NLP, per lo più basate sul linguaggio Python, si allunga di giorno in giorno su GitHub (https://github.com), dove nel maggio del 2021 risultano presenti più di 95.000 *repository*.

answering, l'estrazione di informazione per la costruzione di basi di conoscenza, gli assistenti virtuali, il riassunto o la parafrasi di articoli, la generazione automatica di testi di vario genere, dipende in modo molto significativo dalla qualità dei modelli linguistici generali sulla base dei quali gli algoritmi neurali fanno il loro lavoro.

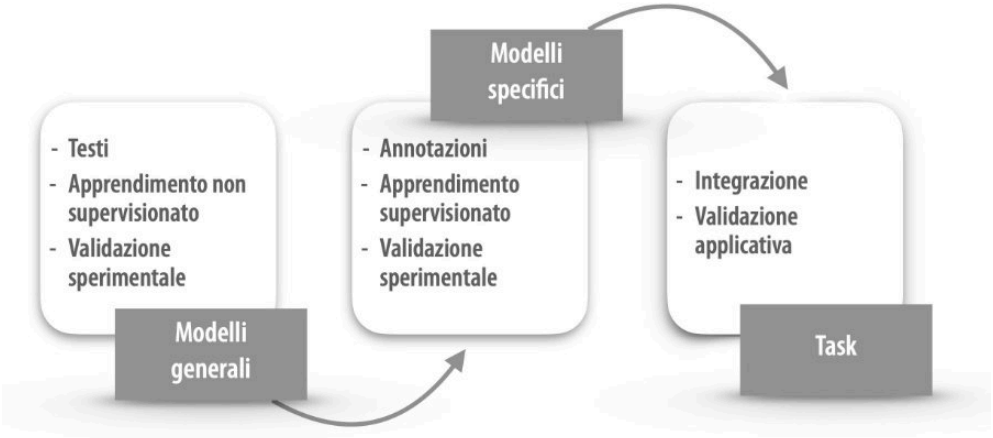


Figura 3: Ciclo di costruzione delle tecnologie linguistiche

Per le aziende e le organizzazioni in possesso di dati, risorse computazionali e competenze interne, tali modelli possono essere prodotti dalle funzioni interne e la qualità necessaria agli scopi applicativi può essere misurata e migliorata in cicli successivi di sviluppo. Ma la situazione è diversa per quello che riguarda le piccole e medie organizzazioni. Poiché la costruzione di modelli linguistici generali mediante le tecniche descritte nella sezione precedente è molto onerosa, il presupposto per l'applicazione delle tecnologie di NLP da parte di aziende medie e piccole diviene la disponibilità di modelli già pronti all'uso (*out-of the box*) disponibili nell'*open source*[10]. La numerosità e la qualità di tali risorse appaiono, tuttavia, fortemente sbilanciate in favore dell'inglese (Tab. 1).

Nei tempi più recenti si assiste ad un incremento dell'offerta di servizi *software* di trattamento del linguaggio naturale (*software-as-a-service*), anche e soprattutto basate sulle infrastrutture *cloud* delle grandi aziende statunitensi[11]. Questi servizi facilitano in modo significativo l'adozione delle

[10] Una vasta comunità produce e rende disponibili tali risorse sulla piattaforma Huggingface (https://huggingface.co/).

[11] Grandi multinazionali come Google, Amazon, Microsoft e IBM ma anche aziende specializzate come Expert.ai offrono oggi, sui loro *cloud*, servizi di NLP.

tecnologie linguistiche, sia per la loro facilità di impiego attraverso interfacce applicative (*application programming interface*), sia per le modalità di imputazione dei costi che in genere sono calcolati in base all'uso (*pay per use*), rendendo così economico l'accesso a tali funzionalità. Alcuni servizi, ad esempio quelli di traduzione automatica[12], sono offerti gratuitamente, in quanto generano, per chi li offre, il vantaggio di raccogliere dati con cui migliorare la qualità dei modelli neurali, da far poi valere all'interno di prodotti commerciali.

La confluenza delle tecnologie del linguaggio nelle grandi piattaforme monopolistiche presenta, per le economie europee, molte criticità, sia di carattere industriale, sia più in generale di carattere socioculturale. Le prime riguardano la dipendenza delle economie dalle tecnologie dei giganti dell'industria informatica statunitense, le seconde sono legate non solo ai temi della protezione dei dati personali[13], ma anche al rischio di "eterodirezione" degli orientamenti e alla omologazione dei modelli comportamentali. Le recenti proposte di regolamento sull'intelligenza artificiale[14] e sui servizi orientati ai dati[15], testimoniano una crescente consapevolezza delle istituzioni europee rispetto a grandi temi geopolitici portati dal progresso tecnologico. Le tecnologie linguistiche sono pienamente coinvolte in queste riflessioni, e anzi ne rappresentano uno snodo fondamentale. Le tecnologie del linguaggio naturale rappresentano infatti uno dei principali settori dell'IA, non solo per la loro immensa potenzialità economica, ma anche per il fatto di affrontare direttamente la materia dell'intelletto umano.

3. Il *knowledge divide* tecno-linguistico

Il linguaggio è il modo in cui rappresentiamo nella coscienza le entità del mondo e costruiamo oggetti sociali[16]. Il fatto che le tecnologie linguistiche privilegino l'inglese rispetto alle altre può dunque amplificare in modo significativo non solo l'anglicizzazione da tempo in atto negli usi specialistici e comuni della lingua (Fisher, Pulaczewska 2008) ma più in

[12] Tra tutti basti citare Google (https://translate.google.com/).

[13] La materia è regolata dalla *General Data Protection Regulation* (GDPR 2016).

[14] Proposta di regolamento del Parlamento europeo del 21 aprile 2012 (Legge sull'intelligenza artificiale).

[15] Commissione europea, Legge sui servizi digitali (https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:52020PC0825) (data ultima consultazione: 29/09/2021).

[16] Nel senso delineato nella teoria del costruzionismo sociale (Searle 1995).

genere i processi di omologazione culturale che accompagnano la globalizzazione. Oltre all'impoverimento della diversità culturale, l'accumulo di conoscenze e capacità linguistiche nelle piattaforme dei monopolisti digitali configura quel *knowledge-divide* che l'UNESCO identificava già nel 2005 come uno dei maggiori rischi a cui andavano incontro le società dell'epoca digitale[17]. Ci si muove infatti nella direzione di una differenza sempre più marcata tra Paesi in grado di gestire e governare la complessità delle tecnologie intelligenti e Paesi che non avranno altra scelta che approvvigionarsene, collocandosi così in una situazione di dipendenza.

Spark NLP models & pipelines	en	es	fr	de	it	*
Named Entities	38	8	5	5	5	
Text Classification	15	0	0	0	0	
Sentiment Analysis	5	0	0	0	0	
Translation	3	0	(1)	(1)	0	
Question Answering	5	0	0	0	0	
Summarization	4	0	0	0	0	
Sentence Detection	1	0	0	0	0	9
Embeddings	101	0	0	0	0	10
Part-of-speech	7	5	2	3	3	
Lemmatization	19	6	3	3	3	
Relation Extraction	12	0	0	0	0	
Spell check	4	0	0	0	1	
TOTAL	**214**	**19**	**11**	**12**	**12**	**19**

Tabella 1: Modelli disponibili in Spark NLP 3.1
(la colonna marcata con asterisco si riferisce ai modelli multilinguistici)

Per quanto concerne le risorse linguistico-computazionali, si può osservare come il numero di modelli disponibili per l'inglese sia di un ordine di grandezza superiore rispetto alla somma di quelli delle altre principali lingue europee[18]. Analizzando, ad esempio, una delle più diffuse piattaforme *open source*[19], si osserva che molti *task* sono realizzabili solo in inglese (Tab. 1).

[17] "*UNESCO World Report: Toward Knowledge Societies*" (UNESCO 2005).

[18] Si sono prese in considerazione le lingue delle maggiori aree economiche.

[19] Spark NLP (https://www.johnsnowlabs.com/) è una delle piattaforme *open source* più diffuse per applicazioni anche di tipo industriale.

Questa disparità si traduce in maggiori capacità di realizzare soluzioni applicative ad alto contenuto di innovazione destinate al mondo anglofono.

La relazione tra la disponibilità di tecnologie linguistiche e l'uso dei mezzi digitali è da dimostrare, tuttavia appare verosimile che l'investimento in tecnologie linguistiche sia legato alla dimensione economica dei mercati interessati alle lingue supportate da quelle tecnologie. Guardando ai dati di EUROSTAT[20], si osserva come l'uso dei media digitali, sia per le imprese, sia per i consumatori, sia generalmente maggiore nei Paesi dove l'inglese è la lingua madre o è molto diffuso nella popolazione. Migliori sono le tecnologie, maggiore sarà la capacità attrattiva delle lingue per le quali sono disponibili; maggiore è la capacità attrattiva di una lingua, maggiori saranno gli investimenti tecnologici ad essa dedicati. Al limite, nella misura in cui l'uso della lingua si dispiega attraverso i mezzi digitali, questa dinamica potrebbe causare l'aumento dell'influenza dell'inglese non solo come interlingua degli affari, ma come lingua di uso comune.

Al di là di qualsiasi giudizio di valore, l'anglicizzazione della vita linguistica europea si presenta come un processo di riduzione della diversità culturale (Fischer, Pulaczewska 2009). Se è vero che le tecnologie linguistiche possono essere, generalmente, parte in causa in questo processo, è anche vero che la traduzione automatica può invece rappresentare un valido strumento di difesa della diversità linguistica. Il celebre aforisma di Umberto Eco (2003): "La lingua dell'Europa è la traduzione" si può oggi generalizzare su scala globale. Quella che al tempo in cui Eco scriveva poteva sembrare una provocazione intellettuale (Eco 2003), ha la possibilità di concretizzarsi grazie alle tecnologie linguistiche. Miliardi di persone connesse attraverso le reti sociali già oggi possono interagire ciascuna usando la propria lingua madre e confidando nella traduzione automatica di ciò che scrivono e che leggono. La qualità di questi servizi, tuttavia, varia molto in funzione delle lingue coinvolte, con l'inglese ancora una volta in posizione dominante. Anche in questo caso, il vantaggio dell'inglese ha origine nella maggiore disponibilità di dati relativi a questa lingua presenti nei *corpora* paralleli necessari all'addestramento delle reti neurali che governano le traduzioni[21]. L'uso dell'inglese come lingua *pivot* è di fatto

[20] EUROSTAT: Digital economy and society, https://ec.europa.eu/eurostat/web/digital-economy-and-society (data ultima consultazione: 29/09/2021).

[21] Si vedano ad esempio le statistiche fornite da European Commission's Directorate-General for Translation: https://wt-public.emm4u.eu/Resources/DGT-TM_Statistics.pdf (data ultima consultazione: 29/09/2021).

l'unica opzione quando si tratta di lavorare con lingue dotate di scarse risorse (Dabre *et alii* 2020). La traduzione come lingua planetaria vedrebbe dunque l'inglese come "garante semantico", come "banca mondiale del dicibile". Una prospettiva di questo tipo non può non sollevare allarme tra chi consideri il linguaggio come principale depositario della vitale specificità delle culture umane. Il passaggio continuo tra *parole* e *langue*[22], la prima espressione della creatività individuale anche estemporanea, la seconda sistema della regolarità intersoggettiva, transiterebbe all'interno di un linguaggio particolare dotato della sua specifica struttura semantica. Questa semantica, per forza di cose, imporrebbe globalmente le concezioni tipiche della cultura nella quale prende corpo.

Una riflessione sul dominio dell'inglese e come questo possa tradursi in un vantaggio geopolitico nella società della conoscenza procede già da alcuni anni (Mendieta *et alii* 2014). Le tecnologie linguistiche possono accentuare o attenuare questo vantaggio: le comunità di ricerca e di pratica possono indirizzare gli sforzi nella direzione del riequilibrio oppure dell'accentuazione delle disparità. La grande industria informatica monopolista statunitense, che investe risorse ingentissime nelle tecnologie linguistiche per accrescere i propri profitti, benché contribuisca in modo considerevole al *software open source* e rilasci dati aperti[23], non si configura oggi come possibile garante di riequilibrio dei rapporti di forza indotti dalle tecnologie. Pertanto, appaiono necessarie specifiche politiche industriali e di ricerca da parte dei sistemi geopolitici non-anglofoni, in particolare quelli europei, finalizzate da un lato alla tutela dell'identità culturale, dall'altro al mantenimento dell'autonomia tecnologica.

4. Cosa fa l'Europa

La *Multilingual Europe Technology Alliance* (META), fondata nel 2010, è una rete che

> riunisce ricercatori, fornitori di tecnologie commerciali, utenti privati e aziendali di tecnologie linguistiche, professionisti del settore linguistico e altre parti interessate della società dell'informazione [in] uno sforzo internazionale ambizioso e congiunto per promuovere la tecnologia linguistica come mezzo per realizzare la visione di un'Europa unita come un unico mercato digitale e spazio informativo[24].

[22] Si tratta di una delle nozioni centrali del *Cours de linguistique générale* di Ferdinand de Saussure (1916).

[23] Fra tutti segnaliamo ad esempio quelli rilasciati da Facebook (https://ai.facebook.com).

Nel corso dell'ultimo decennio, questa iniziativa di coordinamento della ricerca e dell'industria europea si è concretizzata in una piattaforma per la distribuzione di tecnologie e risorse linguistiche (*European Language Grid*, ELG) e nel finanziamento di un progetto (*European Language Equality*, ELE) che ha lo scopo di porre *"le basi per un'agenda strategica e una tabella di marcia per rendere l'uguaglianza del linguaggio digitale una realtà in Europa entro il 2030"*[25].

Nel 2017, il Parlamento europeo ha pubblicato uno studio nel quale si afferma che

> l'emergere di nuovi approcci tecnologici come le reti neurali di deep learning, basati su una maggiore potenza di calcolo e sull'accesso a quantità considerevoli di dati, fa sì che le Human Language Technologies (HLT) diventino una efficace soluzione per superare le barriere linguistiche. Tuttavia, diversi fattori, come la frammentazione del mercato, la mancanza di coordinamento nella ricerca e l'insufficienza dei finanziamenti, ostacolano l'industria europea delle HLT, mettendo le lingue meno dotate di risorse a rischio di estinzione digitale[26].

Si può dunque dire che in Europa vi sia da tempo piena consapevolezza del problema. Tuttavia, considerando l'uscita della Gran Bretagna dall'Unione, il rapporto tra le risorse linguistiche accessibili in ELG e la rilevanza economica (misurata in PIL) delle maggiori lingue europee risulta, ancora oggi, notevolmente sbilanciato in favore dell'inglese. Il programma europeo dovrà dunque essere, nei prossimi dieci anni, estremamente incisivo (Fig. 4).

EUROPEAN LANGUAGE GRID

Lingua	Corpora	Lexica
en	1535	464
es	229	298
fr	250	144
de	295	121
it	139	68

Lingua	GDP 2020
en	306
es	1058
fr	2061
de	2832
it	1479

[24] META, Informazioni generali: http://www.meta-net.eu/meta/about-it (data ultima consultazione: 20/09/2021).

[25] Un punto di partenza per esplorare lo spazio di queste iniziative può essere la pagina iniziale della rete META (http://www.meta-net.eu/).

[26] "*Language equality in the digital age*". https://op.europa.eu/en/publication-detail/-/publication/fa0a50e7-cda4-11e7-a5d5-01aa75ed71a1 (data ultima consultazione: 29/09/2021).

La politica europea nel campo delle tecnologie linguistiche confluisce oggi nel filone più generale della *governance* dell'intelligenza artificiale. L'*Artificial Intelligence Act* proposto dalla Commissione europea, del 21 aprile 2021[27], prevede, tra le varie misure di tutela, stringenti protocolli di verifica delle inferenze condotte dagli automi, in particolare della loro trasparenza e intelligibilità (*auditability*), misure che, applicate ai sistemi linguistici neurali, comportano la necessità di entrare nel merito di come modelli del linguaggio vengono prodotti e utilizzati. In particolare, essendo ricavati da *corpora* testuali non supervisionati e di provenienza eterogenea, i modelli linguistici neurali sono notoriamente esposti al problema dei pregiudizi (*bias*) di varia natura che dai testi vi si riversano. La prevenzione o la rimozione di tali pregiudizi, che rientra senz'altro negli obiettivi delle politiche europee, è tuttavia un problema molto aperto sia sotto il profilo tecnologico, sia sotto quello metodologico (Bolukbasi *et alii* 2016).

Se è vero che la statistica distribuzionale ha dominato lo scenario della ricerca e dell'industria delle tecnologie linguistiche dell'ultimo decennio, superando per molti aspetti i metodi razionali che avevano caratterizzato l'IA "classica", è altrettanto vero che i limiti di questi approcci sono stati già da tempo messi in rilievo (Marcus 2019). Alcune iniziative di ricerca si stanno così orientando verso il recupero della rappresentazione della conoscenza linguistica, a beneficio della trasparenza e della spiegabilità del ragionamento semantico.

5. Prospettive di ricerca

Il peccato originale dei modelli linguistici neurali non supervisionati si può rintracciare nella "scissione dell'atomo semantico" (Vetere 2021). Senza tenere in conto le parole nella loro integrità di segni, l'IA non è verosimilmente in grado di raggiungere una piena padronanza del linguaggio. Con artefatti della dimensione e complessità di GPT-3 si è probabilmente raggiunto il limite di ciò che si può fare con metodi di carattere statistico. Successive versioni della rete neurale, ancora più grandi e costose, potranno migliorare per alcuni aspetti, ma non valicare il confine segnato

[27] Proposta di regolamento del parlamento europeo e del consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'unione, Bruxelles, 21.4.2021, https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:52021PC0206 (data ultima consultazione: 20/09/2021).

dall'approccio che adottano: quello di non perseguire alcuna effettiva comprensione del linguaggio, intendendo per "effettiva comprensione" il passaggio consapevole dal piano del significante a quello del significato.

L'IA classica, sviluppata negli ultimi decenni del Novecento, con le sue "reti semantiche", muoveva da una prospettiva diametralmente opposta rispetto alle reti neurali: quella di codificare uno ad uno i sensi e interpretare i testi alla luce di tali esplicite ipotesi. Risorse come WordNet, VerbNet e FrameNet si inscrivono in quella tradizione e formano la base di approcci semantici ai *task* di NLP (Shi, Mihalcea 2005). Benché accusati di essere troppo difficili da costruire, questi modelli sono di fatto ancora di largo impiego, soprattutto in ambiti specialistici, come ad esempio in medicina[28], dove è importante il controllo razionale della terminologia e dove il rapporto tra dimensioni del lessico e disponibilità di *corpora* settoriali rende difficile l'applicazione di metodi distribuzionali. Molta ricerca si indirizza oggi verso l'uso di modelli concettuali di nuova generazione (detti "ontologie") e la loro integrazione con le potenti tecnologie di apprendimento automatico, proprio per supplire all'insipienza semantica delle reti neurali.

Recenti ricerche hanno mostrato che da risorse lessicali come WordNet si possono ottenere rappresentazioni vettoriali dei vocaboli di una lingua funzionalmente equivalenti, per molti importanti *task*, rispetto agli *embedding* neurali (Saedi *et alii* 2018, Jimenez *et alii* 2019). In sostanza, il metodo consiste nell'estrarre tali rappresentazioni dalla topologia del grafo formato dai nodi delle singole accezioni (sensi) e gli archi delle relazioni lessicali (iponimia, sinonimia, meronimia, etc.) che le connettono. La medesima tecnica è peraltro usata nella estrazione di *embedding* da *knowledge graph* della più varia natura (Wang *et alii* 2017). Si prospetta così una convergenza tra la rappresentazione della conoscenza generale, come quella disponibile nel web sotto forma di *linked open data*[29], e quella contenuta nelle risorse lessicali sotto forma di individuazioni semantiche e relazioni concettuali. Tale convergenza, tuttavia, richiede una riflessione di carattere formale sul rapporto tra lessico e ontologia, la quale sollecita anche alcuni problemi aperti di filosofia del linguaggio. Si tratta di un campo di

[28] Notevole per dimensioni è la terminologia medica multilinguistica di SNOMED (https://www.snomed.org/).

[29] Per una introduzione ai *Linked Open Data* (LOD) si faccia riferimento a https://www.w3.org/wiki/LinkedData (data ultima consultazione: 29/09/2021).

ricerche multidisciplinari che può contribuire in modo significativo al conseguimento degli obiettivi di *governance* delle tecnologie intelligenti che l'Unione europea ha iniziato con decisione a perseguire.

Lo sviluppo di tecnologie per le lingue europee diverse dall'inglese può dunque passare non solo attraverso l'incremento delle risorse con cui alimentare processi di apprendimento non supervisionato e della "forza bruta" computazionale necessaria per supportarli, ma anche attraverso la costruzione di *knowledge graph* linguistici e la loro disponibilità come risorse di conoscenza aperta e verificabile. Quest'ultima si presenta come una prospettiva molto più consona della prima rispetto alle esigenze di una società, come quella europea, alla ricerca di un modello di sviluppo equilibrato e governabile dell'economia digitale.

Conclusioni

Negli ultimi anni, l'intelligenza artificiale basata sulle reti neurali "profonde" ha rivoluzionato le tecnologie linguistiche portando quest'ultime a progredire in numerosi compiti, tra cui, notevolmente, la traduzione automatica. Ma le tecnologie neurali, basate su grandi quantità di dati e ingenti risorse computazionali, aumentano le concentrazioni e il divario in favore delle economie dei Paesi anglofoni che dispongono di maggiori dati linguistici, risorse e competenze specialistiche. Questo divario, se non colmato, può contribuire all'aumento della pressione dell'inglese sulla vita linguistica dei Paesi europei, nonché ad una dipendenza tecnologica in uno dei settori trainanti dell'economia digitale.

Nel corso dell'ultimo decennio, l'Unione europea ha intrapreso azioni per dare impulso all'industria delle tecnologie linguistiche, facendo affidamento sulle sinergie della ricerca pubblica e privata e sulla costruzione di un mercato europeo delle risorse e del *software*. Il raggiungimento di una condizione di equilibrio appare tuttavia ancora lontano. Il regolamento dell'uso delle tecnologie di intelligenza artificiale prospettato recentemente dalla Commissione europea rende ancora più marcata l'esigenza di rafforzare il mercato europeo e renderlo autonomo dai monopoli tecnologici dei Paesi anglofoni, monopoli che le tecnologie linguistiche tendono a rafforzare. I dispositivi di questo regolamento, quando dovesse entrare in vigore, richiederanno ingenti investimenti e dunque un salto di qualità nelle politiche di sostegno nel settore in cui le tecnologie linguistiche si collocano.

Piuttosto che competere con il sistema monopolistico anglo-centrico sul terreno della "forza bruta" dei dati e della computazione, l'Europa può considerare il sostegno alla ricerca di modelli alternativi, più favorevoli ad uno sviluppo decentralizzato, sostenibile e democratico delle tecnologie intelligenti. In questa prospettiva, i risultati ottenuti recentemente mediante risorse che rappresentano in modo trasparente e intelligibile le conoscenze linguistiche appaiono molto incoraggianti.

Bibliografia

Bolukbasi Tolga, Chang Kai-Wei, Zou James, Saligrama Venkatesh, Kalai Adam (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems* 29 (NIPS 2016), https://proceedings.neurips.cc/paper/2016

Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared D, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan Tom, Child Rewon, Ramesh Aditya, Ziegler Daniel, Wu Jeffrey, Winter Clemens, Hesse Chris, Chen Mark, Sigler Eric, Litwin Mateusz, Gray Scott, Chess Benjamin, Clark Jack, Berner Christopher, McCandlish Sam, Radford Alec, Sutskever Ilya, Amodei Dario (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* 33 (NIPS 2020), https://proceedings.neurips.cc/paper/2020

Dabre Raj, Chu Chenhui, Kunchukuttan Anoop (2020). "A Survey of Multilingual Neural Machine Translation". In: *ACM Computing Surveys*, 53-5, 1-38.

Eco Umberto (2003). *Dire quasi la stessa cosa*. Milano: Bompiani.

Goodfellow Ian, Bengio Yoshua, Courville Aaron (2016). *Deep Learning*. Cambridge: MIT Press.

Harris Zelling (1954). "Distributional structure". *Word*, 10/23, 146-162.

Jimenez Sergio, Gonzalez Fabio A., Gelbukh Alexander, Duenas George (2019). "word2set: WordNet-Based Word Representation Rivaling Neural Word Embedding for Lexical Similarity and Sentiment Analysis". *IEEE Computational Intelligence Magazine*, 14/2, 41-53.

Manning Christopher, Schütze Hinrich (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Marcus Gary, Davis Ernst (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon.

Mendieta Eduardo, Phillipson Robert, Skutnabb-Kangas Tove (2014). "English in the geopolitics of knowledge". *Revista Canaria de Estudios Ingleses*, 53, 15-26.

Ferrone Lorenzo, Zanzotto Fabio M. (2020). "Symbolic, Distributed, and Distributional Representations for Natural Language Processing in the Era of Deep Learning: A Survey". *Front. Robot. AI* 6:153. doi: 10.3389/frobt.2019.00153.

Fischer Roswita, Pulaczewska Hanna (eds) (2008). *Anglicisms in Europe: Linguistic Diversity in a Global Context*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Saedi Chakaveh, Branco António, Rodrigues João António, Silva João Ricardo (2018). "WordNet Embeddings". In: *Proceedings of the 3rd Workshop on Representation Learning for NLP Melbourne, Australia, July 20, 2018*, Stroudsburg: ACL, 122-131.

Shi Lei, Mihalcea Rada (2005). "Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing". In: *Computational Linguistics and Intelligent Text Processing*. Berlino Heidelberg: Spring Eds.

Searle John (1995). *The Construction of Social Reality.* New York: Free Press.

Vetere Guido (2021). "Textnology". In: *The Wealth of Languages,* Imminent Research Report 2021, Lebanon Yunction, KY (US), 21-25.

Wang Quan, Mao Zhendong, Wang Bin, Guo Li (2017). "Knowledge Graph Embedding: A Survey of Approaches and Applications". *IEEE Transactions on Knowledge and Data Engineering*, 29/12, 2724-2743, 1 Dec. 2017, doi: 10.1109/TKDE.2017.2754499

Études de cas

De Europa Première partie - Special Issue 2022

Enabling additional official languages in the EU for 2025 with language-centred Artificial Intelligence

Kepa Sarasola, Itziar Aldabe, Nora Aranberri

Introduction

We are in times of change. A technological revolution is taking place right here and right now. With the quality that machine translation applications, voice synthesis and voice recognition have achieved today, we can predict that there will be substantial changes in social communication and in the areas where each language will be used in just a few years.

The futuristic scenario set out in Antoni Olivé's novel *Qui vol el Panglòs* 30 years ago is close to becoming a reality (Olivé 2015). The novel is a reflection on what could happen in a society with an unstable linguistic balance (like Catalonia) if a device—Panglòs—were invented that would enable people to understand any language and to communicate with everyone else using only their own language. In 2021, Panglòs is not a reality. There is still a margin of error when using machine translation, whose size depends on the language pair involved. Consequently, human post-editing is still necessary, but human translators tend to achieve much higher efficiency when using this tool. However, it is highly likely that machine translation technologies will make for smoother flow of knowledge between cultures and among the scientific community, and as a result the pace of social and scientific progress will increase dramatically in the short term.

What is important to note at this stage is that these technologies are not applied only to widely spoken languages; recently, research is also being conducted for low-resourced languages with promising results. Taking Basque as example, the following pages will (1) present instances where new technological tools are proving beneficial for under-resourced languages; (2) describe current research and development projects that afford a glimpse of the next scientific breakthroughs; and (3) put forward

Kepa Sarasola, University of the Basque Country, kepa.sarasola@ehu.eus
Itziar Aldabe, University of the Basque Country, itziar.aldabe@ehu.eus
Nora Aranberri, University of the Basque Country, nora.aranberri@ehu.eus

ideas for taking full advantage of technological advances which would promote under-resourced languages in Europe.

The experience we describe here demonstrates that the new technological developments can consolidate the use of under-resourced languages across Europe. Moreover, it shows that these technologies would allow for a dynamic and affordable provision of language services that could vary according to demand, thus making it possible to recognise all languages as official.

1. A leap forward in language technology

Natural Language Processing (NLP) and Machine Translation (MT) technologies, which rely on Artificial Intelligence, are the heart of today's information processing software. They draw on large amounts of structured and unstructured data collected from texts, as well as from websites and social networks. NLP and MT processors make it possible to analyse and exploit texts in domains such as education, health, law, tourism or the marketplace. Language processors currently offer a long list of applications such as language checkers, language learning systems, machine translation, intelligent search, named entity detection (proper names, dates, brands, products, etc.), document classification, document clustering, document filtering, automatic summary creation, data extraction from documents (sentiment analysis and opinion mining), reputation tracking and monitoring in social networks, alert generation, document queries, chatbots, etc[1].

1.1 Language technology as a tool for language revitalisation

Since 1968, Basque has been engaged in a revitalisation process. Though it has faced formidable obstacles, significant progress has been made in many areas. Six main factors explain the process's relative success: 1) the implementation and official acceptance of Standard Basque[2] (also called "*Batua*") in 1968, 2) integration of Basque in the education system, 3) creation of media in Basque (radio, newspapers, and television); 4) the new legal framework, 5) collaboration between public institutions and people's organisations, and 6) campaigns for Basque language

[1] Plan for the Advancement of Language Technology. Mineco Spanish Government. https://plantl.mineco.gob.es/tecnologias lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan (last access: 31 August 2021).

[2] https://en.wikipedia.org/wiki/Standard_Basque (last access: 31 August 2021).

literacy (Agirrezabal 2010). While these six factors have influenced the revitalisation process, the extensive development and use of language technologies has also had a significant effect (Alegria, Sarasola 2017).

Adding to the thorough work done for Basque since the early days of language technologies, it is expected that the major technological progress of the last six years, such as big data or neural networks, will propel further advances in the language's recovery. This new upsurge may be qualitatively greater for our society. Significant improvements in speech processing, machine translation and text analysis could make a major contribution to facilitating the use of Basque. Needless to say, technological development did not occur without effort. NLP research for Basque has been continuous and active, mainly promoted by a local research group (IXA group, University of the Basque Country). Much work has gone into keeping abreast of the latest research trends and to disseminating the results obtained for Basque in key conferences and international forums. Naturally, this was supported by master- and doctoral-level programmes on NLP.

The Basque language is currently one of the pioneers in methodologies for the promotion of minority languages. This can be seen from the scientific congresses in the field or in the origin of the students enrolled in the master's programme[3] offered by our university. If we make good use of the opportunities offered by technology, in the medium term our programme will also be an international benchmark, with solutions not only for Basque, but also for different languages in international forums and in a multilingual market.

1.2 Rapid increase in LT development

The last six years have seen remarkable developments in language technology. For example, the first scientific publication using deep learning in machine translation appeared in 2014, authored by Bahdanau *et alii* (2014). One year later, 90% of the MT systems winning research challenges were neural systems. The breakthrough in 2015 was the use of Attention-based NMT systems. In 2017, using the Transformer architecture in neural networks brought further improvements. While initial work was carried out for English, given the resources it provides for experimentation, it took just one year to successfully implement a system for Basque, and by 2019, there were not one but five successful systems for Basque.

[3] http://ixa.si.ehu.es/master/ (last access: 31 August 2021).

Currently, the machine translation research community is identifying techniques and strategies for working with languages with limited resources. For example, several efforts have focused on using multilingual data to enrich the tools' knowledge of low-resourced languages (Fan *et alii* 2021). Building domain-specific tools that are also scalable for industry is another key goal.

Translation technology has taken giant steps forward, opening new challenges for bringing multilingual services to our society. We estimate that the production of translated documentation could grow ten-fold in a few years without increasing the number of professional translators.

1.3 Powerful new tools

Over the last three years, several language applications have emerged in the Basque technological landscape that could prove extraordinary catalysts for promoting the use of the language in the public sphere. These applications allow speakers of other languages to understand text and speech in Basque, and Basque speakers to understand texts in other languages. Some of the most noteworthy applications include:

- Elia.eus[4] (Fig. 1), itzuli+[5], and batua.eus[6] machine translators. Similar in operation to the well-known Google Translate, these are three locally developed neural systems that provide high quality translations.
- Content Translation tool[7] allows Wikipedia editors to create translations right next to the original article and automates the boring steps: copying text across browser tabs, looking for corresponding wiki-links, wiki-categories, wiki-templates and programmed components etc. The intrinsic multilingualism of Wikipedia and Wikidata allows, for example, easy translation for all languages of the infoboxes that appear top right in Wikipedia articles. Content Translation offers translation from/into Basque by using elia.eus, Google Translate or Yandex. A first international event[8] was organised in 2021 to allow the research community to take stock of the progress made so far and to identify new avenues for future work.

[4] https://elia.eus/ (last access: 31 August 2021).
[5] https://www.euskadi.eus/itzuliplus/ (last access: 31 August 2021).
[6] https://www.batua.eus/ (last access: 31 August 2021).
[7] https://www.mediawiki.org/wiki/Content_translation (last access: 31 August 2021).

- Aditu[9] bilingual speech recognition (Fig. 2). The Aditu web service recognises both Basque and Spanish speech. It should also recognise English and other languages by 2021. It provides high quality instant transcriptions, automatic generation of subtitles, and direct transcription from the microphone. Anyone can use these applications and then correct transcriptions or subtitles on the online editing interface.
- Interprest[10] interpreting system (Fig. 3). The system's main goal is to offer low-cost and portable interpretation services for different types of event. It is based on mobile phone communication systems, i.e., it is a wireless system. The communication process is simple: the interpreter's mobile phone sends the audio through a small microphone and all attendees can use their own phone to listen to the simultaneous translation. Interprest was a technological platform powered by "San Sebastián 2016", the European Capital of Culture. However, to achieve a low-cost channel forlanguage interpretation in an event, in 2022 another simple and easily accessible possibility is tocreate a group of users in a messaging application (Telegram...) and use only audio in a group call.

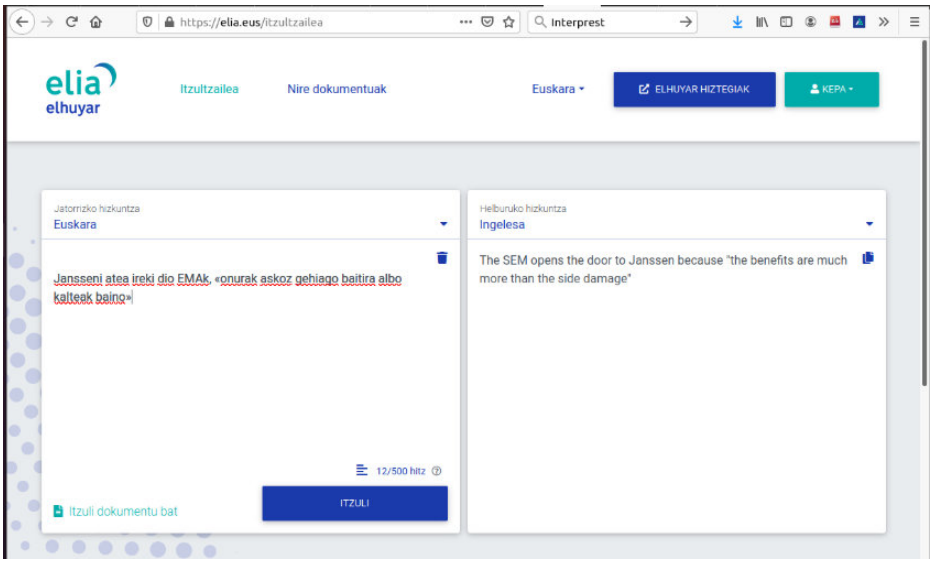


Figure 1: elia.eus machine translator

[8] https://ctn.hkbu.edu.hk/wikiconf2021/ Understanding Wikipedia's Dark Matter. Translation and Multilingual Practice in the World's Largest Online Encyclopaedia (last access: 31 August 2021).
[9] https://aditu.eus/ (last access: 31 August 2021).
[10] https://talaios.coop/2016/09/interprest/ (last access: 31 August 2021).

- Bidaide[11] (Fig. 4) is a web service that allows visitors to a museum, route or building to read or listen to descriptions and general information about the sites on their own mobile phone and in their own language (Cortes *et alii* 2018). Visitors access the information in various ways: by scanning QR codes located in key areas, by GPS positioning (in outdoor routes), or by automatic Bluetooth proximity activation. This makes it accessible even for people with low or no vision. Additionally, this platform also provides the manager of the visited site with advanced language resources to create texts and audios in all relevant languages: machine translation is used to translate texts, while speech synthesis is used to produce audio materials. For accessibility purposes Bidaide uses technologies for speech synthesis and speech recognition. Bidaide uses the visitors' location to guide them along outdoor routes by means of GPS, and indoors through various Bluetooth transmitter beacons. The multilingual content is stored online and is managed by the site team.

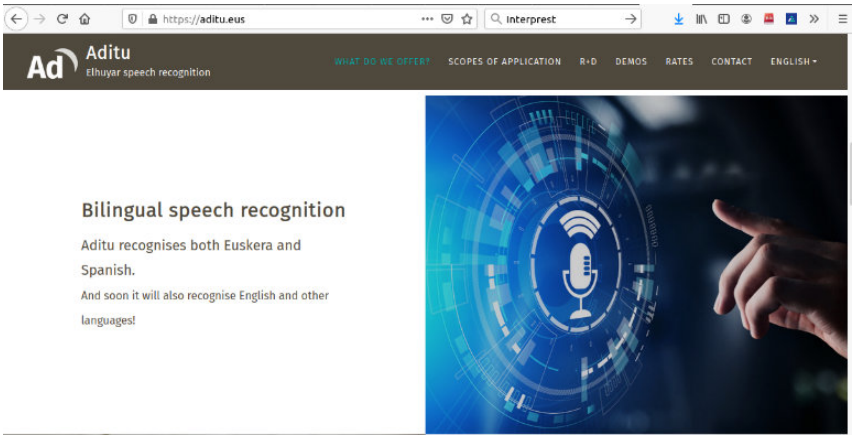


Figure 2: Aditu bilingual speech recognition

Figure 3: Interprest interpreting system

[11] http://bidaide.elhuyar.eus (last access: 31 August 2021).

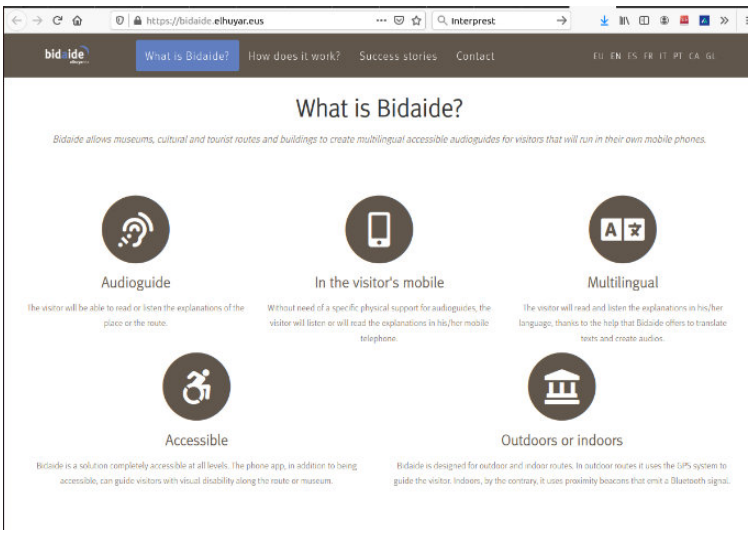


Figure 4: Bidaide, a web service for reading/listening to information on the mobile and in your own language

2. Projects for low-resourced languages

Participation in European initiatives that seek to advance technologies for low-resourced languages has also been key to the digitalisation of Basque. We will briefly describe two such projects, which aim to enhance language technology applications in the European Union: Linguatec and the European Language Equality project.

2.1 Linguatec: cooperation among the languages of the Pyrenees

Linguatec is a project funded by FEDER via POCTEFA (INTERREG V-A Spain-France-Andorra programme). The main objective of Linguatec is to develop, test and disseminate innovative linguistic resources, tools and solutions for a better digitalisation level of the Aragonese, Basque and Occitan languages (Aldabe *et alii* 2019)[12].

The consortium consists of the following partners: 1) Elhuyar Fundazioa (working on Basque and Spanish); 2) Lo Congrès Permanent de la Lenga Occitana (Occitan and French); 3) the University of the Basque Country (Basque and Spanish); 4) CNRS Toulouse Délégation Régionale Midi-Pyrénées (Occitan and French); 5) Euskaltzaindia – Real Academia de la Lengua Vasca (Basque); 6) Sociedad De Promoción y Gestión del Turismo Aragonés (Aragonese).

[12] https://linguatec-poctefa.eu/eu/sarrera/#pll_switcher (last access: 31 August 2021).

In total, the project has developed thirteen main applications since 2018 that facilitate cooperation and interoperability between the languages of the Pyrenees:

- A new translation system for Basque-French.
- A new translation system for French-Occitan.
- A new translation system for Spanish-Aragonese.
- An improved neural translation system for Spanish-Basque.
- A translation app for the languages of the Pyrenees: Basque-French, Basque-Spanish, French-Occitan and Spanish-Aragonese.
- VOTZ, the first tool for speech synthesis in Occitan.
- ReVOc, the first tool for speech recognition in Occitan.
- A Northern Basque speech recognition system.
- New monolingual and bilingual lexicons and morphosyntactic/syntactic analysers for Occitan.
- The Handbook of Unified Basque: "Euskara Eskuz Esku Digitala" by Euskaltzaindia, the Academy of the Basque Language.
- An on-line dictionary of Aragonese, and a roadmap for the Digitalisation of Aragonese.
- A multilingual semantic search engine.
- A system for measuring the vitality of Occitan, Basque and Aragonese.

2.2 ELE: a roadmap to full Digital Language Equality in Europe by 2030

In September 2018, the European Parliament endorsed the report on Language Equality in the Digital Age presented by Jill Evans MEP of Wales with 592 MEPs voting in favour, and with only 45 against and 44 abstentions. Although the report did not become law, it was a declaration made by the European Parliament, which could be used as a reference by all European countries. Until that moment, there were no laws or declarations by the European Parliament to protect low-resourced languages, and decisions regarding their use and promotion thus remained in the hands of the local legislations of each country, which could easily ignore low-resourced languages. The report was a decisive step forward.

Among other things, the report states the following:

> [The EP] Calls on the Commission and the Member States to develop strategies and policy action to facilitate multilingualism in the digital market; requests, in this context, that the Commission and the Member

> States define the minimum language resources that all European languages should possess, such as data sets, lexicons, speech records, translation memories, annotated corpora and encyclopaedic content, in order to prevent digital extinction. (Evans 2018)

This should pave the way to local initiatives for the technological development of minority languages. It remains to be seen what will materialise from this declaration and to what extent real coverage will be given to non-state languages in the coming years.

In this context, the primary goal of the European Language Equality (ELE) project is to prepare the European Language Equality Programme in the form of a strategic research, innovation and implementation agenda and a roadmap for achieving full Digital Language Equality (DLE) in Europe by 2030[13].

Preparing the plan to achieve DLE in Europe by 2030 calls for:

- an accurate and up-to-date description of the 2021 state of technology support for Europe's languages,
- the preliminary definition for achieving full Digital Language Equality in Europe by 2030, and
- identifying gaps and issues regarding LTs, also considering neighbouring disciplines, particularly language-centric artificial intelligence.

The ELE Consortium consists of a total of 53 members: 5 core partners, 9 networks, associations and initiatives, 9 companies and 30 research organisations. In addition to all official European languages, their expertise covers several unofficial, regional and minority languages, either through consortium partners or through the umbrella organisations ELEN and EC-SPM. The consortium as a whole brings together research and industry partners as well as wider networks representing a very broad range of stakeholders that have joined forces to achieve full DLE for all European languages.

3. Additional EU official languages for 2025?

Most countries in the world are multilingual, many officially so. Taiwan, Canada, the Philippines, Belgium, Switzerland, and the European Union are examples of official multilingualism. Under their system, all govern-

[13] European Language Equality (ELE) Project, https://libereurope.eu/project/european-language-equality-ele/ (last access: 31 August 2021).

ment services are available in all of the country's official languages. Also, all citizens may choose their preferred language when conducting business.

In the following paragraphs, we examine some features of practical multilingualism and legal multilingualism in Europe and put forward a number of ideas from the *Global Trends to 2035* report commissioned by the European Parliament (EPRS 2019).

3.1 Practical multilingualism: information in "many" languages

The implementation of practical multilingualism is relative. For example, the website of the European Commission states that the information is provided in the 24 official EU languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish. However, this is not always the case. Indeed, we can read this warning on the website of the Commission[14]:

> We aim to strike a reasonable balance between respect for speakers of the EU's many languages and practical considerations such as limited resources for translation.
> Some content, such as legislation, is always available in all EU languages.
> Other content might be available only in languages that user research tells us will reach the largest audience.
> All content is published in at least English because research has shown that with English we can reach around 90% of visitors to our sites in either their preferred foreign language or their native language.

The criterion applied by the EU for practical multilingualism is clear from the first paragraph: "practical considerations such as limited resources for translation". This is thus one of the key aspects that low-resourced languages, as well as other main languages, must address if all languages are to have the same status.

3.2 Legal multilingualism in Europe: 24 State Languages, tertium genus languages and others

The European Union, with a surface area of more than 4 million square kilometres, is home to more than 60 indigenous languages, which are not

[14] European Commission. Language policy. Information in many languages, https://ec.europa.eu/info/language-policy_en (last access: 31 August 2021).

always confined to being the only language of one of the member states. This diversity, this multilingualism, is one of the characteristics that for centuries have shaped this continent's outlook and essence. The constantly evolving model of the European political structure should continue to be inspired by this intrinsic multilingualism. Until now, however, the officiality of European languages, and hence their compulsory and extensive use in administration and government, has been limited to state languages and has been based on the idea that there is only one language in each member state (state monolingualism). This has shaped the current concept of European linguistic diversity, and as a consequence a hierarchy has been established that prioritises 24 out of the total of 60 European languages (Urrutia 2015).

The idea is that each Member State can establish an official (and working) language. Official languages and working languages are legally equivalent, but in practice English is the most widely used language (French is used by the European Court of Justice in Luxembourg, for historical reasons). The European Agencies have their own language regime.

In principle, Article II-82 of the Treaty establishing a Constitution recognises European linguistic diversity. However, the legal scope of this article is not clear and it is not fully developed; the measures for its application both to state languages and to regional or minority languages have yet to be defined.

The references to the use of languages in the Treaty can be divided into two groups: the references to the rules governing the languages of the EU institutions, and the references to the recognition of European linguistic diversity.

In the first group, the rules governing language arrangements at the EU are based on the concept of constitutional languages, but the Treaty does not define the European official status of each language. In addition, the rules introduce a second level in the constitutional recognition of languages; since some Member States have more than one official language, this second level includes the other languages that enjoy official status in all or part of the Member States' territory (Catalan, Basque, Galician, Breton, Corsican, etc.). The right to petition, however, cannot be exercised in these languages, and citizens cannot demand that they be used officially at any level of the European administration. They are a *tertium genus*, an intermediate category between the languages that benefit from the linguistic rights

recognised by the Constitution and those that are not recognised as having any status in the European institutional context. The legal use of this second intermediate category will depend on subsequent regulatory development, i.e., on the status granted to these languages in future reforms of the rules governing the languages of the EU's institutions (Urrutia 2015).

3.3 The Global Trends to 2035 report

In 2019, the European Parliamentary Research Service published a report on geo-politics and international power entitled "Global Trends to 2035". This document, written by Oxford Analytica, provides a general overview. It discusses eight trends, the third of which is the "Industrial and technological revolution". According to the report, the technologies that would bring about such a revolution include "Artificial intelligence and automation". The report states that:

> One of the largest problems Europe will face in the next two decades is that most of the largest tech providers in the world are based in the United States and China, and their dominance in the sector will be consolidated by the shift to AI.

This raises a question that is as important as it is difficult: how are the big multinationals that control the field of artificial intelligence to be handled? And, consequently, in the face of this oligopoly, what policy should guide public administration? In line with the report's suggestions, we believe that—on the scale that concerns us here, of course—a key principle is "to guarantee the public nature of data and resources, as well as to encourage the work of local companies that can facilitate technological sovereignty".

3.4 A broader multilingualism policy?

We have seen that the concepts of practical multilingualism and legal multilingualism in Europe are relative, and that in practice, the reason to offer (or not offer) official information in one of the languages is highly dependent on the cost of translation. Now, however, a constantly evolving Europe and the current technological revolution in machine translation and language technology have given us a chance for a broader policy of multilingualism to capitalise on Europe's linguistic diversity and to encourage the kind of cooperation among local companies that can fa-

cilitate technological sovereignty. Though such a policy has at times been regarded as controversial, and the idea has been rejected in some other areas where it has been proposed, it has also been seen as necessary for the recognition of different groups or to provide countries with an advantage in presenting themselves to outsiders.

All in all, however, there is no doubt that the recent significant advances in language-centred Artificial Intelligence can be put to positive use in promoting under-resourced languages in Europe. In the current technological scenario, the cost of recognising these languages as official could be affordable, especially since the spheres in which each language would be used officially can be adapted dynamically according to demand.

In the current era of digitalisation, web traffic techniques[15] can measure which content is most frequently visited and which is never visited at all. Currently, language officiality is binary: a language is either recognised as official or it is not. Consequently, all official documentation is translated into all official languages, and nothing is translated into non-official languages. As an alternative, officiality could be defined on an analog basis, with intermediate values, or officiality coefficients, between 0 (non-official language) and 1 (official language). Thus, for example, if a language is assigned a coefficient of 0.7, 70% of official EU documentation could be translated into that language. In that way, data for a given month or year on the web traffic achieved by the texts translated by European Union institutions could help to decide dynamically which 70% of the texts should be translated during the next month or year.

Recognising no official status whatsoever for a language and not producing texts on topical issues not only punishes that language, but also is almost a death sentence for it. Future legal frameworks could establish officiality coefficients for each European language. For example, today's official languages could have a coefficient of 1, those of the *tertium genus* (the Member States' other official languages) could have a coefficient of 0.7 and those with very few resources a coefficient of 0.5.

Conclusions

Certainly, diversity is socially beneficial. Language diversity is too. Europe must succeed in managing its heterogeneity; it should set a mile-

[15] https://en.wikipedia.org/wiki/Web_traffic (last access: 31 August 2021).

stone in the integration of its citizens, becoming a global benchmark. So why not grant EU official status to more languages? Is it expensive? In 2021, the costs associated with it are no longer an excuse.

Yes, creating sterile translations would be a waste of money. Creating translated texts that no one will ever read would be a waste of money. Let us consider that first: what should we translate? Translation supply can be adjusted to demand, to the area and to the depth of information required. Translations can be produced semiautomatically, and decisions about what to translate and in which languages can be made automatically using web traffic techniques. Current technology could allow us to do so, and this goal will encourage the work of local companies that can facilitate technological sovereignty, as the "Global Trends to 2035" report suggests.

Again: recognising no official status for a language not only punishes that language, but also saps Europe's great potential in linguistic diversity. Future legal frameworks could establish officiality coefficients for each European language, so that it is no longer necessary to choose between full recognition and zero recognition.

Let us pay close attention to the latest advances in language-centred Artificial Intelligence. Let us learn from successful experiences not only for highly-resourced language, but also for low-resourced languages, as we have described. And let us, as a result, define the roadmap to additional official languages for 2025.

References

All the references last accessed on 31 August 2021.

Agirrezabal Lore (2010). *The basque experience: some keys to language and identity recovery.* Eskoriatza, Gipuzkoa: Garabide Elkartea.

Aldabe Itziar, Aztiria Josu, Beltrán Francho, Bras Myriam, Ceberio Klara, Cortes Itziar, Coyos Jean-Baptiste, Dazeas Benaset, Esher Louise, Labaka Gorka, Leturia Igor, Sarasola Kepa, Séguier Aure, Sibille Jean (2019). "LINGUATEC: Desarrollo de recursos lingüísticos para avanzar en la digitalización de las lenguas de los Pirineos". *Procesamiento del Lenguaje Natural,* 68, 159-162. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/

Alegria Iñaki, Sarasola Kepa (2017). "Language technology for language communities: An overview based on our experience". In: *Mercator-SOAS-CIDLeS & FEL Conference Communities in Control: Learning tools and strategies for multilingual endangered language communities* (CinC 2017), October 19-21, Alcanena: Portugal, 91-97.

Bahdanau Dzmitry, Cho Kyunghyun, Bengio Yoshua (2014). *Machine Translation by Jointly Learning to Align and Translate*. https://arxiv.org/abs/1409.0473, Arxiv.org

Cortes Itziar, Leturia Igor, Alegria Iñaki, Astigarraga Aitzol, Sarasola Kepa, Garaio Manex (2018). "Massively multilingual accessible audioguides via cell phones". EAMT 2018. European Association of Machine Translation. Alicante.

EPRS, European Parliamentary Research Service (2019). *Global Trends to 2035. Geo-politics and international power*. https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2017)603263

Fan Angela, Bhosale Shruti, Schwenk Holger, Ma Zhiyi, El-Kishky Ahmed, Goyal Siddharth, Baines Mandeep, Celebi Onur, Wenzek Guillaume, Chaudhary Vishrav (2021). "Beyond English-centric multilingual machine translation". *Journal of Machine Learning Research*, 22(107), 1–48.

Instituto Cervantes (2015). *Presentación del "Plan de Impulso de las Tecnologías del Lenguaje"*. OCLC 1164865977. https://coleccionesdigitales.cervantes.es/digital/collection/prensa1/id/9634/ Instituto Cervantes, Madrid.

Mineco (2015). *Plan for the Advancement of Language Technology. Spanish Government*. https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan-Advancement-Language-Technology.pdf

Olivé Antoni (2015). Qui vol el panglòs? Barcelona: Voliana Edicions.

Urrutia Iñigo (2015). "Régimen jurídico de las lenguas y reconocimiento de la diversidad lingüística en el Tratado por el que se establece una Constitución para Europa". *Revista de Llengua i Dret*, 42, 231-273.

De Europa Première partie - Special Issue 2022

Langages et savoirs : intelligence artificielle et traduction automatique dans la communication scientifique*

Maria Luisa Villa, Maria Teresa Zanola, Klara Dankova

Introduction

Au cours de ces derniers temps, une discussion s'est déclenchée autour de la construction de systèmes intelligents artificiels qui maîtrisent la morphologie complexe d'une langue ou d'un système de traduction, visant l'efficacité du résultat et non pas nécessairement la fidélité de la simulation de la capacité humaine. L'intelligence artificielle (IA) peut toutefois être utilisée pour modéliser les capacités linguistiques humaines et contribuer à accroître nos connaissances sur les objets, se prêtant à devenir un instrument précieux pour mieux comprendre les mécanismes mentaux humains étudiés dans le cadre des neurosciences cognitives. L'IA n'a pas encore passé le stade de la maîtrise absolue de la langue : des algorithmes très différents — suivant ce qu'on leur demande de faire — définissent la façon de penser de l'IA, du connexionnisme au symbolisme. Des données adéquates au sujet traité sont indispensables pour atteindre la modélisation recherchée : il n'est pas possible de capturer toute la réalité possible dans un corpus, aussi vaste qu'il soit. Les formes d'annotations — entités, relations entre elles, chaînes de coréférences — gagnent de plus en plus d'importance, pour se soumettre à l'exploration sémantique des données ; ces données sont traitées, corrigées, nettoyées des « bruits », découpées, enrichies de métadonnées… La méthodologie de travail intervient ainsi, de manière à pouvoir traiter, entre autres, les mots composés, l'écriture inclusive, les conjugaisons des verbes, évaluer

Maria Luisa Villa, Università Statale di Milano et Accademia della Crusca, marialuisavilla40@gmail.com
Maria Teresa Zanola, Università Cattolica del Sacro Cuore, mariateresa.zanola@unicatt.it
Klara Dankova, Università Cattolica del Sacro Cuore, klara.dankova@unicatt.it
*Maria Luisa Villa est l'auteure du paragraphe 1, Maria Teresa Zanola de l'introduction, du paragraphe 2 et de la conclusion ; Klara Dankova est l'auteure de l'Annexe 1.

les informations extralinguistiques, les néologismes. L'IA entre en dialogue avec les connaissances, avec les savoir-faire les plus variés, pour nous offrir un instrument qui nous dominera… ou qui restera assujetti à nos volontés et à notre guide.

Nous allons proposer quelques réflexions sur ce sujet de deux perspectives différentes, scientifique et linguistique, afin d'offrir un cadre de référence sur des problématiques aussi passionnantes et surprenantes, et des suggestions sur le rapport entre langages et savoirs à l'aune de l'IA, considérant le rôle de la traduction automatique dans la communication scientifique. Après une analyse du rôle du langage dans la communication scientifique, et du dépassement de l'emploi d'une seule langue internationale vers les différentes langues nationales grâce à la diffusion de la traduction automatique, l'attention est focalisée sur l'importance de la terminologie dans la communication scientifique et sur l'intérêt de son utilisation correcte au sein des traitements numériques.

1. Les sciences et la barrière linguistique : la *lingua franca* comme remède

Le langage de la science ne naît pas spontanément, mais doit être produit au prix de nombreux efforts. Afin d'accueillir ses concepts, la langue doit élaborer un vocabulaire spécifique en adaptant les mots et les phrases à ses besoins épistémiques particuliers. Étudier la science, c'est étudier un nouveau langage où des mots nouveaux — tel que « quark » l'a été au moment de sa apparition — ou des mots de tous les jours — tels que « cellule », « noyau », « énergie », « travail » — acquièrent une valeur sémantique déterminée (en physique), différente de leurs sens habituels.

Dans les sciences, les mots usuels deviennent des termes ayant une signification définie de manière univoque et universellement partagée. Cependant, chaque langue a sa propre structure et son histoire : dans le domaine des sciences, la transmission des idées véhiculées dans « une » langue scientifique, parmi des locuteurs qui ont des origines linguistiques différentes, pourrait devenir problématique et faire surgir des évocations sémantiques à partir des langues sources respectives. Cette barrière a été atténuée au fil du temps en attribuant à une langue, appelée « *lingua franca* » (Brosch 2015), la suprématie en tant qu'instrument de communication internationale entre des personnes de langues maternelles différentes, pour lesquelles cette langue est une langue étrangère. Une *lingua*

franca est alors une langue délibérément acquise pour des raisons d'utilité afin de faciliter la communication entre des locuteurs de langues maternelles différentes (Villa 2016 : 128-131).

Trois facteurs principaux ont favorisé la transformation d'une langue en *lingua franca* : la création d'un empire, le commerce et les religions. Cependant, les langues *lingua franca* ne durent pas indéfiniment : la disparition des causes historiques, militaires, commerciales ou religieuses qui ont décrété leur prédominance, finit par décider leur déclin.

Le latin a été pendant quelques siècles la langue de la science, la *lingua franca* dans laquelle s'est exprimée la révolution scientifique de l'époque baroque. Galilée, Kepler et Newton ont écrit en latin. Newton a intitulé *Philosophiae Naturalis Principia Mathematica* l'ouvrage dans lequel il décrit la loi de la gravitation universelle (1687). Le XVIII^{e} siècle, qui est le siècle de la rationalisation du langage scientifique, est aussi l'époque de l'abandon du latin, qui survit en tant que matrice de la nomenclature spécialisée et de la terminologie, mais qui est de moins en moins utilisé en tant que langue de communication. La recherche scientifique s'est adressée aux langues nationales, la science s'est exprimée en français, anglais, allemand, italien et suédois.

Les élites académiques d'Europe ont toutefois conservé une connaissance assez solide du latin pendant au moins deux siècles. Le langage scientifique, bâti sur ce noyau terminologique commun, semblait pouvoir s'exprimer dans tous les langages scientifiques de l'époque avec la même adéquation et la même précision, offrant un répertoire ouvert à la compréhension mutuelle.

1.1 La multiplicité des langues : un danger épistémique ?

L'illusion de l'universalité s'est dissoute au cours du XX^{e} siècle, lorsque l'affirmation de la civilisation industrielle a augmenté le nombre de savants, étendu les tâches de la science et déplacé ses frontières bien au-delà de l'Europe. La science est née avec ses projets à grande échelle, ses énormes financements, ses équipements complexes et ses laboratoires internationaux.

Le poids des retombées technologiques a transformé les règles de diffusion et d'application des connaissances scientifiques : en changeant le monde, la science a aussi changé les conditions de son propre développement.

Face à l'énorme richesse des connaissances générées par une collaboration de plus en plus internationale, les chercheurs ont commencé à percevoir la multiplicité des langues comme un danger pour les normes épistémiques de la « science ouverte ». Ceux qui publient dans une langue de diffusion « inférieure » par rapport aux langues les plus utilisées risquent de briser l'unité du système scientifique mondial, car ils soustraient leurs résultats à l'examen des pairs (Villa 2018). Ensuite, l'histoire politique et militaire de l'Occident a imposé sa direction et l'anglais est devenu la langue véhiculaire de la science. En quelques décennies, déclarait *The Economist* en décembre 1996, « l'anglais a acquis une position inattaquable en tant que langue standard du monde : il est devenu une partie intrinsèque de la révolution mondiale des communications » (*The Economist*, 21 décembre 1996 : 39).

À l'époque de l'économie des connaissances, l'anglais n'est pas seulement le noyau technique du langage scientifique, mais aussi la langue utilisée pour parler de science et pour tous les débats scientifiques.

La richesse des connaissances accumulées au cours de décennies d'utilisation a transformé l'anglais dans un outil précieux pour la communication internationale de la science. Les profonds changements intervenus dans les équilibres politiques et économiques pourraient favoriser la diffusion concurrentielle d'autres langues, mais jusqu'à présent, aucune ne semble avoir la force de remplacer l'anglais en tant que *lingua franca* principale de la science (Gordin 2015).

L'anglais scientifique semble représenter ainsi un héritage potentiellement durable : les risques pour l'anglais ne viennent donc pas de l'histoire, mais du progrès technologique. En 2010, un livre très bien documenté et inattendu (Ostler 2010) avançait le fait que les *linguae francae* seraient supplantées non pas par des événements militaires, commerciaux ou religieux, mais par des innovations technologiques. Les progrès rapides de la traduction automatique rendraient toute *lingua franca* inutile : l'anglais deviendrait ainsi la dernière *lingua franca* de l'histoire de l'humanité.

Dans le modèle d'Ostler, l'avenir serait multilingue à un degré beaucoup plus radical que le passé, mais uniquement parce que personne n'aurait le besoin pratique d'apprendre des langues étrangères. Lorsque l'anglais universel sera devenu un simple souvenir, chacun utilisera les mots dans la langue qui lui convient davantage, qui est la plus facile pour lui, sans avoir à se soucier de la langue de ses auditeurs. Le monde sera compréhensible dans toute sa diversité.

1.2 La révolution technologique et l'IA : de la maîtrise du sens à la sémantique distributionnelle

Le traitement automatique du langage, généralement appelé par l'acronyme anglais NLP (*Natural Language Processing*), est l'une des branches les plus complexes du domaine de l'IA (Chiari 2007). Depuis quelques années, les informaticiens tentent une approche rationaliste du langage, dans l'illusion d'instruire les machines à comprendre la sémantique des mots. Ce projet s'est avéré improductif et a été remplacé par une approche empirique qui ne tient pas compte de la compréhension des machines et qui est connue sous le nom de sémantique distributionnelle. C'est le système qui semble bien fonctionner dans de nombreuses applications.

Les progrès récents de l'apprentissage automatique avec les réseaux neuronaux profonds ont suggéré que le problème pourrait être résolu en fournissant aux machines des volumes de données suffisants —les téraoctets. La méthode NPL a été la clé du succès des projets d'IA consacrés au langage : c'est grâce à ces systèmes qu'un progrès remarquable a été réalisé dans le domaine de la traduction automatique au cours des dernières années.

La méthode NPL neuronale est basée sur une stratégie d'évitement du problème du sens des mots : elle adopte l'hypothèse distributionnelle et affirme qu'il existe une relation entre le sens des mots et la façon dont ils sont distribués — c'est-à-dire, qu'ils sont nécessaires, qu'ils sont utilisés et qu'ils se combinent — dans les textes. Puisque des mots de sens similaire apparaissent dans les mêmes contextes, on peut supposer que la sémantique distributionnelle peut servir de base à l'application de techniques d'apprentissage automatique. Le réseau nous fournit des téraoctets de textes et il devient donc relativement facile et mathématiquement aisé de rapprocher le sens d'un mot de sa distribution.

La sémantique distributionnelle a une base empirique qui fonctionne bien dans de nombreuses applications : elle permet d'évaluer la similarité de sens entre les mots et les phrases avec plus de précision, de force et de flexibilité que ce que l'on pourrait faire en utilisant des vocabulaires de synonymes ou d'autres ressources lexicographiques. Cette capacité est cruciale pour de nombreuses applications d'apprentissage automatique, telles que les classificateurs de texte ou les traducteurs automatiques.

Pour ce type d'applications, il n'est pas nécessaire de connaître la signification des mots : il suffit de disposer d'une mesure de leur similarité due au fait qu'ils apparaissent dans les mêmes phrases (par exemple, « caresser

le chien/ le chat »). Grâce à une grande quantité de données valables et aux ressources nécessaires, les matrices de distribution fonctionnent aujourd'hui mieux que n'importe quel dictionnaire dans de nombreux cas pratiques.

1.3 Quelques nouvelles utiles sur le portable : « traduire le site web »

Les progrès de la traduction automatique sont désormais dans l'expérience de tous. De nombreux téléphones portables permettent de lire un texte scientifique ou un journal dans plusieurs langues autres que l'originale. Malheureusement, toutes les langues ne sont pas représentées, et souvent l'italien figure parmi celles-ci. Une nouvelle de très grande portée pour la communication scientifique italienne est apparue aux premiers mois de 2020 : le lancement de *Nature Italy*, un supplément en ligne de la revue *Nature*[1].

Ci-dessous le texte de l'un des premiers numéros, consacré à la lutte contre la pandémie de coronavirus dans les laboratoires italiens. Il suffit de cliquer sur le lien « lire en italien » et le texte apparaît immédiatement.

Il s'agit d'une petite et grande révolution : la langue italienne paraît là où elle n'avait pas le droit de parole. La communauté scientifique, qui sait mêler réalisme solide et utopie inébranlable, est le creuset où les forces internes qui favorisent le développement de la connaissance peuvent s'intégrer aux forces externes, qui définissent son destin institutionnel. Le libre recours à la langue maternelle, en plus de l'utilisation d'une langue internationale et véhiculaire, ouvre de nouveaux espaces de liberté (Villa 2018 : 150).

Figure 1 : La capture d'écran de l'un des premiers numéros de *Nature Italy* (21.11.2020)

[1] Voir dans l'Annexe 1 les analyses développées sur les traductions automatiques à partir aussi de la revue *Nature*.

2. Langues et savoirs, langages et terminologies

La langue est un moteur de l'évolution scientifique : au cours des siècles, le rôle de la terminologie est devenu de plus en plus visible, constituant un pont entre les connaissances, un tissu de toutes les connexions conceptuelles, visuelles et matérielles. En organisant le savoir par secteur d'activité et en l'identifiant linguistiquement, la terminologie a fait de la langue un instrument de progrès.

L'idée de créer un rapport solide entre la théorie et le langage avait permis à Lavoisier de suivre la *Logique* de Condillac et d'en appliquer les indications précieuses : grâce au lien étroit entre les faits, les mots et les idées, le langage devint le lieu de la rupture avec la tradition antérieure et le lieu où l'on défendait la vérité, au lieu de transmettre les erreurs et les préjugés. L'expérience de la nomenclature de la chimie du XVIII^{e} siècle (Villa 2016 : 125-128) reste un modèle encore aujourd'hui : puisque le langage et la connaissance sont inséparables, seul le langage peut structurer et organiser l'information acquise par le sens ; en refaisant le langage, on peut refaire la science. La communication de la nouvelle science chimique commence donc par la nouvelle formulation de la terminologie, où les termes nouvellement créés sont incorporés dans la textualité écrite et orale : la réforme systématique du langage de la chimie, qui rejette tout fondement naturel et historique des termes à utiliser, met en évidence le rôle de la terminologie, dont la construction fait partie intégrante de la connaissance.

Une dimension culturelle accompagne ainsi ce processus descriptif et cognitif, dans lequel la terminologie permet de définir l'histoire, l'identité et la conceptualité du domaine concerné. La terminologie devient le moteur de cette évolution ; elle favorise la diffusion de l'innovation, en la greffant sur le patrimoine linguistique antérieur et en l'insérant dans un réseau conceptuel précis.

Cette entreprise organisationnelle et normative, fondée sur l'idée d'une relation dialogique entre la science/ les sciences et le langage, établit un lien presque organique entre la chose et le mot, partant tantôt de la continuité, tantôt de l'arbitraire. Face à la complexité du réel et à la richesse foisonnante de sa terminologie technique et scientifique — qui a pour tâche d'inventorier, de classer et de structurer certaines nomenclatures (il suffirait de penser aux terminologies des arts et métiers) —, des constellations de termes s'assemblent et prennent forme, telle une création microcosmique, dont l'ambition atteint les ontologies et les liens sé-

mantiques du réseau. L'objectif est non seulement de rationaliser et de pérenniser le savoir-faire au sein d'une discipline ou d'un métier, mais surtout de faciliter la communication entre les usagers et les professionnels ainsi que de rendre ce savoir accessible à tous (Zanola 2020 : 69-70).

La terminologie ne doit pas générer de confusion dénominative, qui peut être nuisible et contre-productive pour les sciences. La terminologie se prête à devenir une occasion de communiquer de nouveaux concepts dans une langue donnée, elle permet une diffusion plus claire et plus précise de l'ensemble des termes et des concepts d'une discipline.

Les innovateurs scientifiques et techniques sont les architectes créateurs de la terminologie : la création d'un lexique spécialisé devient l'un des moyens de transférer la nouvelle technique, les nouvelles connaissances scientifiques. L'attention est portée non seulement sur l'acte de nommer le concept ou l'objet, mais en même temps sur la définition du terme : la définition complète la systématisation du nouveau terme qui s'insère dans le champ considéré, en trouvant son propre espace sémantique à l'intérieur du réseau conceptuel dans lequel le nouveau terme arrive.

La description systématique des sciences et des techniques — de la même manière que celle des arts et des métiers — met en place une procédure qui ne cessera jamais d'être une préoccupation constante de chaque langue et de chaque institution chargée de sa promotion et de sa diffusion. Une terminologie précise est alors une valeur et un atout, qui favorise la rencontre entre les approches linguistico-culturelles et technoscientifiques, et qui permet de formuler et de résoudre les problèmes liés à la traduction multilingue de concepts véhiculés par des termes spécialisés dans les langues naturelles.

La terminologie est le patrimoine et la richesse expressive de chacun et offre les outils de ses activités aux spécialistes et aux traducteurs, aux journalistes scientifiques et aux rédacteurs techniques, à toutes les catégories professionnelles ainsi qu'au citoyen ordinaire dans chaque métier et profession, scientifique et humaniste, institutionnel et juridique, économique et financier, technique et opérationnel. C'est le véhicule qui transmet et exprime toute nouveauté dans le patrimoine conceptuel (Zanola 2018).

Plus la terminologie est consciente de ses composantes linguistiques, conceptuelles et culturelles, plus le degré d'efficacité communicative est

élevé. Comment préserver ce patrimoine de langages et savoirs dans la communication scientifique avec la traduction automatique ?

Nous allons proposer quelques réflexions sur les avantages et les inconvénients de la traduction automatique par rapport au traitement terminologique, en perspective plurilingue.

2.1 Terminologie et traduction automatique : c'est moins mécanique qu'on ne le pense...

La première étape pour la qualité d'une communication au citoyen et au grand public est sans aucun doute l'utilisation correcte de la terminologie et la diffusion d'une terminologie claire et précise, ce qui constitue une valeur économique et juridique en soi. La terminologie et son utilisation correcte peuvent également constituer un instrument de diffusion des objets et des concepts, contribuer à en accroître la diffusion et pouvoir être ainsi acquis dans l'établissement des grands corpus qui serviront à « nourrir » les ressources utiles pour la traduction automatique neuronale. Il ne suffit pas seulement de pouvoir disposer de banques de données rigoureuses, mais aussi de pouvoir comprendre leur distribution par rapport aux types de textes en général et de textes scientifiques hautement spécialisés ou destinés à la vulgarisation.

Les technologies de traduction automatique (TA) se sont considérablement améliorées au cours des deux dernières décennies, avec des développements dans la TA statistique basée sur les phrases (SMT) et, récemment, dans la TA neuronale qui permet d'améliorer l'opérabilité et la qualité de la pratique de la traduction, de répondre exigences de l'industrie de la langue, d'accélérer les communications entre différents pays et différentes langues...

Si l'on considère les inconvénients liés à la traduction automatique, nous nous limitons à renvoyer aux analyses et aux commentaires qui prennent en compte les facteurs suivants : le manque de précision terminologique et les fautes linguistiques, relatives aux calques fautifs, aux nuances aspectuelles et à l'incohérence temporelle, aux changements dans l'ordre des mots, aux transformations structurelles — telles que la suppression d'arguments, l'adjonction d'arguments, l'inversion des relations de dépendance —, à l'effacement d'informations, aux problèmes de référents et aux biais de genre. Les irrégularités terminologiques sont elles aussi à signaler (Schumacher 2019 : 120). Pour ce qui concerne les

avantages, les facteurs gagnants sont plutôt d'ordre stratégique : la possibilité d'avoir accès à l'information scientifique dans plusieurs langues, la valorisation de toutes les langues, l'évidence du fait que ce qui a été réalisé pour la communication technique peut atteindre aussi la communication scientifique (cf. Boitet 2007 ; Grass 2010 ; Cronin 2013 ; Martikainen, Kübler 2016 ; Schumacher 2019).

La plupart de ces considérations reposent sur la disponibilité de grandes données parallèles pour l'entraînement des systèmes de TA, des ressources qui ne sont pas disponibles pour la majorité des langues ; par conséquent, les technologies actuelles ne sont souvent pas en mesure d'être appliquées aux langues à faibles ressources (Liu *et alii* 2020).

La TA est désormais utilisée pour surmonter les barrières linguistiques dans les milieux à haut risque tels que les hôpitaux et les tribunaux, mais la recherche sur son utilisation dans les domaines de la santé et du droit peut souvent ignorer les complexités de la langue et de la traduction effectuée par des humains (Nunes Vieira *et alii* 2021). La connaissance des atouts spécifiques et, surtout, des limites de la TA devient alors indispensable. Dans son état actuel, la technologie de la TA peut exacerber les inégalités sociales et mettre certaines communautés d'utilisateurs en plus grand danger : c'est une question qui mérite une attention plus forte de la part des chercheurs et des décideurs politiques.

Les avancements des travaux (Popel *et alii* 2020) montrent aussi que la TA peut se rapprocher de la qualité de la traduction humaine et la dépasser en termes d'adéquation dans certaines circonstances, ce qui amène à considérer que la TA a le potentiel de remplacer les humains dans les applications où la conservation du sens est l'objectif principal.

2.2 La traduction automatique et le plurilinguisme

La TA peut ainsi améliorer l'opérabilité, la qualité et la rapidité de la traduction et répondre aux exigences de la pratique de la traduction, accélérant la communication entre différents pays et différentes langues. Il s'agit de garantir la sauvegarde des connaissances et de protéger le patrimoine intellectuel représenté par la terminologie dans les langues sources.

La terminologie est un facteur de protection de la communication publique et institutionnelle, un facteur de cohésion sociale et de sauvegarde de la diversité culturelle et linguistique. La relation claire entre la nomen-

clature et l'objet et/ ou le concept est la garantie d'une communication correcte, respectueuse, compétente et experte, et favorise une communication professionnelle de qualité.

Les différentes manières d'utiliser la TA, le recours au traitement massif des données (*big data*) se mesurent sur le terrain avec les compétences des traducteurs, qui doivent de plus en plus travailler pour vérifier et améliorer la traduction disponible selon les possibilités de la TA. Les traducteurs et les traductrices, les interprètes vont ainsi réorienter leurs compétences professionnelles, leur productivité, en fonction de la portion de données que leur langue de travail peut offrir.

Conclusion

L'accord est désormais unanime sur l'importance du plurilinguisme et sur la protection des diversités linguistiques. Stickel (2005) avait souligné que la traduction automatique, grâce au recours à des banques de données terminologiques, pourrait permettre de préserver la communication dans les langues spécialisées dans chaque langue.

Cette perspective est largement partagée : il faut aussi reconnaître les efforts d'évolution dans la formation des professionnels de la traduction, pour lesquels on peut bien observer les attentes de plus en plus exigeantes par rapport au marché de travail. L'enjeu est d'ailleurs des plus importants : faire vivre une langue, c'est permettre qu'elle se manifeste, qu'elle se transmette, qu'elle s'épanouisse. L'altérité, la diversité a sa place dans cette liberté des langues et de leur expression.

> Être reconnu, c'est occuper une place dans l'esprit d'un ou de plusieurs autres et, d'une manière plus générale, dans la société. Il n'y a pas là seulement une question de justice ou d'injustice, mais, plus radicalement, l'enjeu d'être ou de ne pas être (Flahault 2009 : 461).

Annexe 1

L'utilisation de la traduction automatique dans la communication multilingue - Fiche de synthèse

1. Groupe de recherche

Chiara Arrigoni, Federica Ciurlia, Maria Cristina Denaro, Luca Ghidini, Lucrezia Marzo, Nicolò Pulici, Anna Lisa Rossi, Sofia Vigo.
Cours de Terminologies et Politiques Linguistiques Master 1,
Prof. Maria Teresa Zanola dir., *Università Cattolica del Sacro Cuore*, Milan.
Coordination : Klara Dankova.

2. Présentation

Dans le cadre des études les plus récentes sur l'utilisation de l'intelligence artificielle dans le domaine de la traduction, ce projet de recherche vise à évaluer l'utilisation de la traduction automatique dans la presse spécialisée et généraliste dans une tranche temporelle contemporaine (mars-avril 2021). Avec un groupe d'étudiants de Master 1 et 2, spécialisés dans différentes langues et cultures, le projet a développé des analyses détaillées considérant une perspective multilingue ; plus précisément, les textes — en original ou en traduction — qui ont été examinés concernent les langues suivantes : anglais, italien, français, allemand, espagnol et portugais.

2.1 Objectifs

Le projet vise à vérifier si les articles traduits en différentes langues sur les sites web d'un certain nombre de périodiques ont fait l'objet d'une révision humaine. Les analyses effectuées contribuent à détecter les principales différences entre la traduction automatique et la traduction avec intervention humaine.

2.2 Méthodologie

Chaque article a été analysé à partir du texte dans la langue originale, de sa traduction officielle proposée sur le site web de la revue et de la traduction automatique générée par le système DeepL. Les erreurs et les écarts générés par la traduction automatique ont été détectés et recensés selon une analyse qui s'étend sur trois niveaux :

- morphologique : genre, nombre, forme verbale ;
- lexical : choix lexicaux, choix terminologiques, traitement des expressions lexicalisées, sigles et acronymes, métaphores ;
- syntaxique : emploi des collocations, ordre des mots, emploi des prépositions, emploi de l'article.

Les éléments observés dans la traduction automatique ont été comparés avec ceux de la traduction publiée dans le site web du périodique, afin de vérifier la qualité de la traduction officielle et celle de la traduction automatique par rapport au texte source.

2.3 Le corpus d'analyse

Le corpus est constitué de 25 articles disponibles dans la langue originale et dans une autre langue (traduction officielle) qui ont été collectés sur les sites web des périodiques respectifs, et de leurs traductions fournies par DeepL pendant la période du 22 mars au 9 avril 2021. La taille des articles originaux est de 139 663 caractères (espaces comprises), ce qui correspond, à peu près, à 77,5 pages standard de 1 800 caractères (espaces comprises). Du point de vue de la typologie textuelle, les articles analysés peuvent être distingués en trois catégories :

- articles publiés dans des revues de vulgarisation scientifique : *National Geographic* (3 articles), *Nature Africa* (4 articles), *Nature Italy* (9 articles), *Journal de la Haute Horlogerie* (3 articles) ;
- articles de la presse quotidienne : *New York Times* (5 articles) ;
- un article de la presse populaire : *Egoista* (1 article).

Les analyses ont été effectuées en tenant compte des combinaisons de langues, réparties en trois groupes : traductions de l'anglais (EN -> X), traductions vers l'anglais (X -> EN) et traductions de et vers des langues autres que l'anglais (X -> X). Le tableau suivant résume le nombre d'articles, leur provenance et leur longueur (le nombre de caractères – espaces comprises – du texte original) pour chaque combinaison linguistique :

Combinaison linguistique		Nombre d'articles	Sources des articles	Nombre de caractères
EN -> X	EN -> IT	10	Nature Italy (9), New York Times (1)	52.742
	EN -> DE	4	National Geographic (3), New York Times (1)	39.344
	EN -> FR	4	Nature Africa	9.866
	EN -> ES	1	New York Times	5.790
X -> EN	DE -> EN	1	New York Times	7.187
	FR -> EN	3	Journal de la Haute Horlogerie	12.582
	PT -> EN	1	Egoista	5.942
X -> X	IT -> ES	1	New York Times	6.210

Tableau 1 : La taille du corpus et les combinaisons linguistiques considérées

3. Exemples d'analyse

Les différences principales entre les traductions automatiques et celles qui sont réalisées avec l'intervention humaine peuvent être illustrées à partir de quelques exemples tirés du corpus d'articles écrits en anglais et traduits en italien (par. 3.1) et en allemand (par. 3.2)[2] :

3.1 La combinaison linguistique EN -> IT

a) niveau morphologique				
	f	revue EN	revue IT	DeepL (IT)
Nombre	7	at the universities of Groningen and Rome	alle università di Groningen e Roma	all'università di Groningen e Roma
Forme verbale	5	The light emitted by a country at night is an excellent index of GDP, but it does not mean that I earn more money by turning on the light.	La luce emessa da un paese di notte è un ottimo indicatore del PIL, ma questo non vuol dire che si guadagni di più accendendo la luce.	La luce emessa da un paese di notte è un ottimo indice del PIL, ma non significa che io guadagni di più accendendo la luce.
Genre	3	[…] it is harder for female professors to […], despite being just as productive […].	[…] è più difficile per le professoresse […], nonostante siano altrettanto produttive […].	[…] è più difficile per le professoresse […], nonostante siano altrettanto produttivi […].

Tableau 2 : Exemples d'erreurs au niveau morphologique (EN -> IT)

b) niveau lexical				
	f	revue EN	revue IT	DeepL (IT)
Choix lexicaux	142	Governments beyond Italy are now in danger of following the same path, repeating familiar mistakes and inviting similar calamity.	I governi d'oltralpe rischiano ora di seguire la stessa strada, reiterando errori noti e ripetendo disastri simili.	I governi oltre l'Italia sono ora in pericolo di seguire lo stesso percorso, ripetendo errori familiari e invitando simili calamità.
Choix terminologiques	25	[…] by making them more vulnerable to windstorms, fires, and insect outbreaks, […].	[…] rendendole più vulnerabili a tempeste di vento, incendi e attacchi degli insetti.	[…] rendendole più vulnerabili alle tempeste di vento, agli incendi e alle epidemie di insetti, […].
Expressions lexicalisées	8	But tracing the record of their actions shows missed opportunities and critical missteps.	Ma andando a ripercorrere le loro azioni si possono notare alcune opportunità mancate e critici passi falsi.	Ma tracciare il registro delle loro azioni mostra opportunità mancate e passi falsi critici.
Sigles et acronymes	6	The potential role of DCT [digital contact tracing] has been highlighted since the early days of the pandemic.	Il ruolo potenziale del tracciamento digitale è stato evidenziato sin dai primi giorni della pandemia.	Il ruolo potenziale del DCT è stato evidenziato fin dai primi giorni della pandemia.

Tableau 3 : Exemples d'erreurs au niveau lexical (EN -> IT)

[2] 'f' indique la fréquence absolue de chaque type d'erreur détecté.

c) niveau syntaxique				
	f	revue EN	revue IT	DeepL (IT)
Emploi des collocations	20	We can nail down opportunities and risks that are not so evident through traditional analysis.	Riusciamo a individuare opportunità e rischi che non sono così evidenti dalle analisi tradizionali.	Possiamo inchiodare opportunità e rischi che non sono così evidenti attraverso l'analisi tradizionale.
Ordre des mots	26	But tracing the record of their actions shows missed opportunities and critical missteps.	Ma andando a ripercorrere le loro azioni si possono notare alcune opportunità mancate e critici passi falsi.	Ma tracciare il registro delle loro azioni mostra opportunità mancate e passi falsi critici.
Emploi des prépositions	41	Forests in Finland, northern European Russia and the Alps emerged as the most fragile ecosystems, followed by warm–dry forests [...].	Le foreste in Finlandia, nel nord Europa della Russia e nelle Alpi emergono dallo studio come gli ecosistemi più fragili, seguite da foreste calde e secche [...].	Le foreste della Finlandia, della Russia settentrionale europea e delle Alpi sono emerse come gli ecosistemi più fragili, seguite dalle foreste caldo-secche [...].
Emploi de l'article	2	33.4 billion tonnes of biomass (58 per cent of Europe's total forest mass)	33,4 miliardi di tonnellate di biomassa (il 58% della massa forestale totale dell'Europa)	33,4 miliardi di tonnellate di biomassa (58% della massa forestale totale dell'Europa)

Tableau 4 : Exemples d'erreurs au niveau syntaxique (EN -> IT)

3.2 La combinaison linguistique EN -> DE

a) niveau morphologique				
	f	revue EN	revue DE	DeepL (DE)
Forme verbale	4	The damage can still be reversed, he and his colleagues say.	Der Schaden könne immer noch rückgängig gemacht werden, sagen er und seine Kollegen.	Der Schaden kann immer noch rückgängig gemacht werden, sagen er und seine Kollegen.
Genre	3	"Immunity is waning, but certainly not gone, and I think this is key," says Lavine (f), who wasn't involved with the study.	„Die Immunität lässt nach, aber sie verschwindet definitiv nicht, und ich denke, das ist der Schlüssel", sagt Lavine, die nicht an der Studie beteiligt war.	„Die Immunität lässt zwar nach, ist aber noch nicht verschwunden, und ich denke, das ist der Schlüssel", sagt Lavine, der nicht an der Studie beteiligt war.

Tableau 5 : Exemples d'erreurs au niveau morphologique (EN -> DE)

b) niveau lexical				
	f	revue EN	revue DE	DeepL (DE)
Choix lexicaux	38	He is a slight man whose mischievous mien and goatee give him the air of one of Dumas's three musketeers.	Er ist ein eher zierlicher Mann, dessen spitzbübische Miene und Spitzbart ihm den Hauch eines der drei Musketiere von Dumas verleihen.	Er ist ein schmächtiger Mann, dessen spitzbübische Miene und Spitzbart ihm die Ausstrahlung eines der drei Musketiere von Dumas verleihen.
Choix terminologiques	16	About 5.4 million acres burned in 2019, an area roughly the size of New Jersey.	Im Jahr 2019 brannten etwa zwei Millionen Hektar, eine Fläche etwa so groß wie Sachsen-Anhalt.	Im Jahr 2019 brannten etwa 5,4 Millionen Hektar, eine Fläche etwa so groß wie New Jersey.
Expressions lexicalisées	1	This heated exchange over little illustrated several things.	Diese hitzige Debatte über Kleinigkeiten machte mehrere Dinge deutlich.	Dieser hitzige Austausch über wenig illustriert mehrere Dinge.
Métaphores	3	"We're going to be able to manage it because of modern medicine and vaccines, but it's not something that will just vanish out of the window".	„Dank moderner Medizin und Impfstoffen werden wir in der Lage sein, es in den Griff zu bekommen. Aber es ist nichts, das sich einfach in Luft auflösen wird".	"Wir werden in der Lage sein, es dank der modernen Medizin und Impfstoffe in den Griff zu bekommen, aber es ist nicht etwas, das einfach aus dem Fenster verschwinden wird".

Tableau 6 : Exemples d'erreurs au niveau lexical (EN -> DE)

c) niveau syntaxique				
	F	revue EN	revue DE	DeepL (DE)
Emploi des collocations	5	"There will be pockets of people who won't take [the vaccines] […]."	„Es wird zwar Gruppen von Menschen geben, die [die Impfstoffe] nicht nehmen […]."	„Es wird Taschen von Menschen geben, die [die Impfstoffe] nicht nehmen […]."
Ordre des mots	7	Julien Denormandie, the agriculture minister, called the mayor's embrace of the meatless lunch "shameful from a social point of view" […].	Julien Denormandie, der Landwirtschaftsminister, bezeichnete den vom Bürgermeister verordneten Übergang zu einem fleischlosen Mittagessen als „aus sozialer Sicht beschämend" […].	Julien Denormandie, der Landwirtschaftsminister, nannte die Umarmung des fleischlosen Mittagessens durch Bürgermeister „beschämend aus sozialer Sicht" […].
Emploi des prépositions	19	Julien Denormandie, the agriculture minister, called the mayor's embrace of the meatless lunch "shameful from a social point of view" […].	Julien Denormandie, der Landwirtschaftsminister, bezeichnete den vom Bürgermeister verordneten Übergang zu einem fleischlosen Mittagessen als „aus sozialer Sicht beschämend" […].	Julien Denormandie, der Landwirtschaftsminister, nannte die Umarmung des fleischlosen Mittagessens durch Bürgermeister "beschämend aus sozialer Sicht" […].

Tableau 7 : Exemples d'erreurs au niveau syntaxique (EN -> DE)

4. Résultats de l'analyse

Différents types d'erreurs intervenus dans la traduction automatique (par. 2.2) ont été étudiés afin d'identifier leur pertinence dans l'ensemble

du corpus. Le tableau ci-dessous montre les fréquences absolues (f) de chaque type d'erreur détecté :

Niveau	Type de faute	f	f (total)
a) morphologique	Nombre	15	35
	Forme verbale	14	
	Genre	6	
b) syntaxique	Emploi des prépositions	74	149
	Ordre des mots	46	
	Emploi des collocations	26	
	Emploi de l'article	3	
c) lexical	Choix lexicaux	294	378
	Choix terminologiques	55	
	Expressions lexicalisées	16	
	Sigles et acronymes	8	
	Métaphores	5	

Tableau 8 : Les fréquences absolues des différents types d'erreurs

Un nombre total de 562 erreurs a été identifié et analysé aux niveaux morphologique, syntaxique et lexical. En référence à notre corpus (77,5 pages standard), il est possible d'affirmer qu'une moyenne de 7 erreurs de traduction automatique correspond à une page standard (du texte original). Les erreurs les plus fréquentes appartiennent au niveau lexical (378), suivies des erreurs syntaxiques (149) et morphologiques (35). Concrètement, les trois types d'erreurs les plus fréquentes concernent les choix lexicaux (294), l'emploi des prépositions (74) et les choix terminologiques (55). En ce qui concerne les traductions des articles proposés sur les sites web des revues considérées, la recherche a montré des emplois différents de la traduction automatique selon les cas.

En particulier, dans le cas de *Nature Italy* (revue spécialisée) et du *New York Times* (quotidien), l'utilisation de la traduction automatique s'accompagne de l'intervention humaine qui fait la différence, offrant un style élaboré et de nombreux ajouts efficaces pour la clarté du texte. Les articles en italien de *Nature Italy* se distinguent par une terminologie correcte et cohérente, les auteurs des deux articles (en anglais et en italien) étant de langue maternelle italienne et ayant une bonne connaissance de la terminologie en italien. De même, les traductions du magazine *National Geographic* montrent une utilisation évidente d'un traducteur automatique malgré de nombreux remaniements et des changements stylistiques. La traduction automatique a également été utilisée dans le cas de la revue *Nature Africa*, mais cette fois avec des résultats très différents, qui vont des erreurs évidentes dans la traduction proposée à des cas où la traduction automatique a offert de bons résultats. Finalement, les traductions des articles de la revue spécialisée *Journal de la Haute Horlogerie* et de la revue populaire *Egoista* diffèrent fortement des traductions automatiques, ce qui nous amène à conclure qu'elles ont été effectuées par un traducteur humain qui n'a pas eu recours à la TA.

Remarques finales

La traduction automatique accompagnée d'une intervention humaine joue un rôle crucial dans les traductions des articles de tous les périodiques de notre corpus, à l'exception du *Journal de la Haute Horlogerie* et de *Egoista* (deux périodiques qui sont, sans aucun doute, moins répandus que les autres qui ont été considérés dans le corpus). Le nombre relativement faible d'erreurs détectées (7 erreurs/1 page standard) démontre un bon fonctionnement des traducteurs automatiques. Toutefois, il ne faut pas oublier que les traductions automatiques présentent de nombreuses erreurs stylistiques qui n'ont pas été prises en compte dans l'analyse. À titre d'exemple, nous mentionnons l'expression « *a handful of studies* » (texte original EN), traduite par DeepL au moyen d'un calque « *une poignée d'études* » (FR), au lieu de proposer une expression plus courante, telle que « *quelques études* » (traduction officielle FR). Les analyses ne reflètent pas non plus des traitements inappropriés des références culturelles, comme dans « *eine Fläche etwa so groß wie New Jersey* » (traduction automatique DE), censé d'exprimer « *an area roughly the size of New Jersey* » (texte original EN). Même si le manque d'adaptation ne représente pas, dans le cas présent, une erreur grave, parce que cela n'entraîne aucun changement de sens, une bonne adaptation des références culturelles par un traducteur humain est toujours préférable car elle peut faciliter considérablement la compréhension du message de la part du public visé (par exemple, « *eine Fläche etwa so groß wie Sachsen-Anhalt* », traduction officielle DE).

Pour conclure, la présente recherche a montré que la révision des traductions automatiques par un traducteur-réviseur humain s'avère nécessaire malgré les performances de plus en plus élevées des logiciels de traduction, et cela à tous les niveaux analysés : morphologique, syntaxique et lexical. D'un point de vue quantitatif, l'analyse a révélé que les phénomènes relevant du niveau lexical sont la source la plus importante d'erreurs. Dans le cas des textes de spécialité, la nécessité d'une révision approfondie s'impose notamment sur le plan terminologique. Nos analyses des traductions automatiques ont fait ressortir que dans certains cas, l'utilisation incorrecte ou incohérente des termes peut altérer de manière significative le contenu du message du texte original ; un exemple frappant nous est fourni par le traitement des unités de mesure : par exemple, « 5,4 millions d'acres » (texte original EN) — « *5,4 Millionen Hektar* » (traduction automatique DE) — « *zwei Millionen Hektar* » (traduction officielle DE). Une attention particulière doit également être accordée au traitement des unités terminologiques désignant de nouveaux concepts, car celles-ci ne sont pas encore suffisamment représentées dans les corpus de référence utilisés par les traducteurs automatiques. Dans notre corpus

c'est le cas des termes « *digital contact tracing* » (texte original EN) et « *physical distancing measures* » (texte original EN), liés à la pandémie du Covid-19. La comparaison entre les équivalents en italien utilisés dans les traductions officielles (respectivement « *tracciamento (digitale) dei contatti* » ou « *contact tracing* » et « *distanziamento sociale* ») et les solutions proposées par DeepL (respectivement « *tracciatura digitale dei contatti* » et « *misure di allontanamento fisico* ») ne laisse aucun doute sur l'utilité d'une révision de la part d'un expert humain ayant des connaissances précises sur l'emploi des termes dans la langue cible.

Bibliographie

La date de dernière consultation de tous les liens hypertextes est le 12 septembre 2021.

Boitet Christian (2007). « Corpus pour la TA : types, tailles et problèmes associés, selon leur usage et le type de système ». *Revue française de linguistique appliquée*, XII/1, 25-38.

Brosch Cyril (2015). « On the Conceptual History of the Term Lingua Franca ». *Apples: journal of applied language studies*, 9/1, 71-85.

Chiari Isabella (2007). *Introduzione alla linguistica computazionale*. Bari : Laterza.

Cronin Michael (2013). *Translation in the Digital Age*. Londres-New York : Routledge.

Flahault François (2009). « Reconnaissance et anthropologie générale ». Dans : Alain Caillé, Christian Lazzeri (éds). *La reconnaissance aujourd'hui*. Paris : CNRS Editions, 455-469.

Gordin Michael D. (2015). *Scientific Babel : How Science Was Done Before and After Global English*. Chicago, Illinois : University of Chicago Press.

Grass Thierry (2010). « A quoi sert encore la traduction automatique ? ». *Les Cahiers du GEPE*, 2. URL : <http://www.cahiersdugepe.fr/index.php?id=1367>

Liu Chao-Hong, Karakanta Alina *et alii* (2020). « Introduction to the Special Issue on Machine Translation for Low-Resource Languages ». *Machine Translation*, 34, 247-249.

Martikainen Hanna, Kübler Natalie (2016). « Ergonomie cognitive de la post-édition de traduction automatique : enjeux pour la qualité des traductions », *ILCEA*, 27. URL : <https://journals.openedition.org/ilcea/3863>

Nunes Vieira Lucas, O'Hagan Minako *et alii* (2021). « Understanding the societal impacts of machine translation : a critical review of the literature on medical and legal use cases ». *Information, Communication & Society,* 24/11. DOI: 10.1080/1369118X.2020.1776370

Ostler Nicholas (2010). *The Last Lingua Franca. English Until the Return of Babel*. Londres : Allen Lane.

Popel Martin, Tomkova Marketa *et alii* (2020). « Transforming machine translation : a deep learning system reaches news translation quality comparable to human professionals ». *Nature Communications,* 11. URL : < https://doi.org/10.1038/s41467-020-18073-9>

Schumacher Perrine (2019). « Avantages et limites de la post-édition ». *Traduire*, 241, 108-123. URL : <http://journals.openedition.org/traduire/1887>

Stickel Gerhard (2005). « Plurilinguismus und Übersetzen : Investition in Europas Zukunft ». Dans : Fritz Nies (éd). *Europa denkt mehrsprachig / L'Europe pense en plusieurs langues*. Tübingen : Gunter-Narr-Verlag, 97-106.

The Economist (1996). « Language and Electronics: the coming global tongue ». *The Economist*, 21 décembre 1996, 37-39.

Villa Maria Luisa (2016). *La scienza sa di non sapere per questo funziona*. Milan : Guerini.

Villa Maria Luisa (2018). *Scienza è democrazia. Come funziona il mondo della ricerca*. Milan : Guerini.

Zanola Maria Teresa (2018). *Che cos'è la terminologia*. Rome : Carocci.

Zanola Maria Teresa (2020). « Lingua e linguaggi, un ponte fra arte e scienza ». Dans : Gaspare Polizzi (éd). *Arte & Scienza*. Naples : Doppiavoce, 69-78.

Variation et traduction

De Europa Première partie - Special Issue 2022

Terminologie, intelligence artificielle, psychologie cognitive : réflexions sur les interactions possibles dans l'étude de la variation en langue spécialisée

Anne Condamines

Introduction

Après une période de rapprochement fructueuse en lien avec l'analyse automatique de corpus spécialisés, au début des années 1990, les relations entre l'intelligence artificielle (IA) et la terminologie se sont beaucoup distendues lors des dernières années, avec le développement de la sémantique distributionnelle et de l'apprentissage profond dans le TAL (Traitement Automatique de la Langue). En IA, la construction et la mise à jour des ontologies, apparentées à des réseaux terminologiques, se font désormais, le plus souvent, sans que soit fait appel à des connaissances linguistiques. Dans le même temps, certains psychologues ont cru voir, dans les (parfois très bons) résultats de l'IA obtenus grâce à des méthodes dites de sémantique distributionnelle, une confirmation de leurs hypothèses sur l'apprentissage et la mise en œuvre du lexique mental, qui se feraient indépendamment de situations particulières. Cette position pose question à la fois pour la recherche en terminologie et pour la recherche en psychologie cognitive. D'une part, les études en terminologie, qui, par essence, prennent en compte une situation liée à une connaissance spécialisée, s'intéressent de plus en plus à la notion de variation, en fonction de différents paramètres extralinguistiques. D'autre part, de nombreux chercheurs en psychologie cognitive plaident en faveur d'une cognition incarnée et située. Les relations entre ces trois disciplines sont donc complexes. Tout en tenant compte de cette complexité, cet article s'interroge sur les possibilités de complémentarité entre ces disciplines, avec un point de vue orienté vers les besoins en analyse de la variation dans les corpus spécialisés.

Anne Condamines, CNRS et Université de Toulouse, anne.condamines@univ-tlse2.fr

Il est d'abord fait un état des lieux des relations bilatérales entre les trois disciplines : terminologie et intelligence artificielle, intelligence artificielle et psychologie cognitive, psychologie cognitive et terminologie (1er paragraphe). Dans le paragraphe 2, l'étude d'un phénomène particulier, l'alternance de construction (directe *vs via* une préposition) de *pêcher* et de *rivière*, montre comment la variation peut intervenir dans les discours spécialisés en fonction de l'implication affective du pêcheur, et s'interroge sur les possibilités que ce type de variation soit repéré par des méthodes d'IA.

1. Des relations bilatérales

Cette partie rend compte des relations existant entre les trois disciplines : terminologie, IA et psychologie cognitive afin de poser le cadre de la réflexion. Dans le dernier sous-paragraphe 1.4, nous mettons en question la façon dont l'IA emprunte la notion de « sémantique distributionnelle » à la linguistique pour justifier son approche.

1.1 Terminologie et IA

À la fin des années 1980, la terminologie et l'ingénierie des connaissances (IC) (un des aspects de l'intelligence artificielle) se sont rapprochées sur la base de différents constats.

a) Les deux disciplines utilisaient des modes de représentation de la connaissance constitués de réseaux de concepts reliés par des relations. Termes et relations étaient des mots (voire des groupes de mots le plus souvent).
b) Les deux disciplines se sont orientées vers l'utilisation de corpus pour repérer les termes et les relations. Cela permettait à l'IC de ne pas construire les systèmes à base de connaissances à partir des seuls entretiens avec des experts (pas toujours à l'aise dans cet exercice). Quant à la terminologie, elle pouvait bénéficier des développements de la linguistique de corpus et utiliser des méthodes similaires à celles utilisées pour la lexicographie à base de corpus.
c) Dans les deux cas, les méthodes de construction des réseaux de concepts se basaient sur la mise en œuvre de connaissances linguistiques, en particulier des marqueurs de relations (en lien avec la notion de « *knowledge rich context* » proposée par Meyer (2001). Pour mémoire,

les marqueurs de relations sont des éléments lexico-syntaxiques qui permettent de repérer, plus ou moins systématiquement, des relations en corpus. Par exemple : [Tous les N1 sauf les N2] permet de repérer une relation d'hyperonymie entre N1 et N2.

Exemple : Tous les poissons sauf les truites sont relâchés.

Ce rapprochement de l'IA et de la terminologie s'est manifesté en particulier par la création des « bases de connaissances terminologiques » (Meyer *et alii* 1992 ; Condamines 2018c) et par le développement des études sur les marqueurs de relations conceptuelles et de leur variation en contexte, par exemple en fonction du genre textuel (Auger, Barrière 2008 ; Condamines 2002 ; Marshman *et alii* 2008).

En France, le groupe TIA (Terminologie et Intelligence Artificielle) créé par Didier Bourigault et Anne Condamines en 1993 a permis à des chercheurs de différentes communautés scientifiques : linguistique, terminologie, informatique, sciences de l'information, de se réunir et d'interroger les rapprochements possibles autour du thème général « corpus et terminologie » (Aussenac, Condamines 2007).

Depuis une dizaine d'années, les deux disciplines se sont éloignées. La terminologie, avec le développement de la terminologie textuelle, a évolué vers la prise en compte d'autres objectifs que la seule construction de réseaux de concepts à partir de textes (Condamines 2018 ; Condamines, Picton à paraître). Quant à l'IA, c'est moins la construction de réseaux conceptuels (ontologies) en tant que tels qui devient l'objectif des études, mais plutôt des applications finales, utilisant (ou pas) des ontologies : recherche d'information, traduction, question-réponse, recherche de thématiques… (Turney, Pantel 2010). L'intérêt commun entre les deux communautés (IA et terminologie), en particulier pour la recherche et l'étude des relations conceptuelles, s'est donc largement émoussé et les disciplines se sont éloignées.

1.2 Psychologie cognitive et IA

Le développement des méthodes d'apprentissage du sens mettant en œuvre la « sémantique distributionnelle » *via* la construction de vecteurs en IA (il faut le reconnaître, avec un succès certain du point de vue des applications) a ravivé les discussions entre différents courants de la psychologie cognitive à propos du « lexique mental » (Segui 2015). D'une

part, il a montré l'importance du rôle de l'environnement syntaxique des mots (leur distribution) pour l'élaboration du sens, ce qui n'était pas forcément un aspect reconnu en neuropsychologie. Mais, d'autre part, il a pu laisser penser que l'apprentissage des mots se faisait par la mémorisation de ces seuls contextes langagiers, syntaxiques (collocations), indépendamment des situations dans lesquelles ils apparaissent.

> *Cognitive scientists have argued that there are empirical and theoretical reasons for believing that VSMs [Vector Space Model of Semantics], such as LSA and HAL, are plausible models of some aspects of human cognition* (Turney, Pantel 2010 : 144).

Cette hypothèse est fortement contestée par les tenants de la cognition incarnée (*grounded cognition*) (Barsalou 2003), qui estiment que l'apprentissage des mots ne peut se faire sans la prise en compte de la situation extralinguistique mais aussi des réactions sensori-motrices des locuteurs/ auditeurs.

> *We draw two conclusions [...]. The first is that high-dimensional theories such as LSA and HAL are inadequate accounts of human meaning because the symbols (high dimensional vectors) are not grounded* (Glenberg, Robertson 2000 : 384).

Les méthodes de traitement automatique de la langue inspirées de la sémantique distributionnelle tiennent le plus souvent à l'écart cette « inscription corporelle » de la connaissance, certains auteurs remettant carrément en cause cette dimension.

> *The importance of embodiment and grounding is exaggerated, and the implication that there is no highly abstract representation at all, and that human-like knowledge cannot be learned or represented without human bodies, is very doubtful* (Landauer 1999 : 624).

Cet éloignement est expliqué de la façon suivante par Glenberg et Robertson : « *The reason for using ungrounded symbols is clear : they are far easier to use in computer and mathematical simulations than are grounded representations* » (Glenberg, Robertson 2000 : 399). Toutefois, des études tentent de montrer que les deux approches concernant la connaissance lexicale ne sont pas incompatibles dans la perspective de la sémantique distributionnelle au sein de l'IA ; la prise en compte de la dimension incarnée est ainsi en train de se développer :

> *There is a growing trend in cognitive sciences to find a common ground in which embodied cognition and distributional approaches to meaning could eventually meet* (Lenci 2008 : 25).

1.3 Terminologie et psychologie cognitive

Nous l'avons vu, la réflexion sur les relations entre terminologie et intelligence artificielle s'est développée à partir des années 1990. Pour une part, cet état de fait, qui a encouragé la terminologie à travailler sur des données réelles (des corpus) est lié à une réaction contre la vision prescriptive de la terminologie proposée par Wüster dans la Théorie Générale de la Terminologie à partir des années 1930 (*General Theory of Terminology*, GTT) (Wüster 1976). Avec la GTT il s'agissait d'établir des référentiels terminologiques, sorte de normes permettant une diffusion espérée transparente des informations, entre entreprises d'un domaine, dans une langue ou entre langues. Aussi louable qu'il soit, ce point de vue est peu compatible avec le fonctionnement discursif, qui, par essence, se libère parfois des normes, en fonction du contexte (partage de la connaissance dans des communautés restreintes), en fonction du besoin (efficacité de la communication), de l'évolution dans le temps, de la nécessité de créer des variantes en fonction de l'apparition de nouveaux concepts, etc. Plusieurs auteurs se sont opposés à cette vision prescriptive. Certains se sont inscrits dans la perspective de la linguistique de corpus outillée (la terminologie textuelle par exemple). D'autres dans la perspective de la théorie des *frames* (Faber 2012). D'autres chercheurs se sont appuyés sur la sociologie comme la socioterminologie (Gaudin 2003), d'autres encore à la fois sur la sociologie et la psychologie comme la terminologie sociocognitive (Temmerman 2000). Cette prise en compte des aspects sociologiques et psychologiques du fonctionnement terminologique a marqué le rapprochement avec les travaux en linguistique sociocognitive (Kristiansen, Dirven 2008) et en linguistique de corpus (Gries 2015). La variation des termes, entendus comme des unités (ou des poly-unités) lexicales, ne pouvait donc plus être ignorée.

Quant à la dimension « affective » de la terminologie, elle n'est que très rarement prise en compte, ce que regrettent des auteurs comme Baumann :

> [...] *theories of emotion which have been ignored by LSP research for a long time are of increasing methodological and methodical significance because they offer far-reaching strategic orientations for the communicative-cognitive analysis of information processing in LSP texts*. (Baumann 2007 : 322).

Cette dimension émotionnelle fait écho à la dimension incarnée de la psychologie cognitive.

1.4 Linguistique et IA : la question de la « sémantique distributionnelle »

La plupart des travaux en TAL se revendiquent actuellement d'une approche distributionnelle. Or, dans sa prise en compte par les outils, ce point de vue est en partie amputé des propositions des origines.

Depuis les premières réflexions sur le distributionnalisme dans les années 1930, inspirées par Bloomfield, l'approche distributionnelle en linguistique a beaucoup évolué. À partir des années 1950, deux principaux courants sont apparus, l'un, américain, avec comme chef de file Harris (Harris 1954) et l'autre, anglais, avec comme chef de file Firth (Firth 1957). La méthode harrissienne, inspirée par une vision behavioriste du fonctionnement du sens, préconise une approche mathématique pour arriver, par différents types d'opérations, au sens intrinsèque d'une phrase. L'approche anglaise est beaucoup plus ancrée dans une vision sociologique (voir ci-dessous). C'est d'abord l'approche harrissienne qui a été convoquée en TAL pour justifier les travaux sur le sens en lien avec l'accès à de très gros volumes de données textuelles. Mais, tout l'appareillage mathématique de la méthode harrissienne n'étant pas mis en œuvre dans cette approche en TAL (on a d'ailleurs plutôt parlé d'une méthode « à la Harris »), c'est l'autre courant de l'analyse distributionnelle qui a ensuite été convoqué, celui de Firth. On ne compte ainsi plus le nombre de travaux de TAL en sémantique distributionnelle qui se revendiquent de Firth en citant cet extrait : « *You shall know a word by the company it keeps* » (Firth 1957 : 11). Le problème est que cette phrase est isolée du cadre général de l'approche firthienne. Un autre extrait de Firth est, à cet égard, parlant :

> *First the structure of the appropriate contexts of situation must be stated. Then the syntactical structure of the texts. The criteria of distribution and collocation should then be applied* (Firth 1968 [1952] : 19).

Ainsi, pour Firth (et l'école londonienne), la situation de communication contribue à la construction du sens et joue même un rôle déterminant. Notons d'ailleurs que la situation est aussi prise en compte dans la proposition de Harris, qui a créé le concept de sous-langage, sous-partie de la langue associée à des domaines de connaissances particuliers. Même si les études évoluent, en particulier dans l'analyse de corpus spécialisés (Fabre *et alii* 2014), la situation de communication est assez peu prise en compte dans les travaux en TAL qui s'appuient sur la sémantique distributionnelle. La principale explication est liée au fait que les méthodes d'ap-

prentissage profond se mettent en place sous la forme de constitution de vecteurs (Heylen, Bertels 2016), ce qui nécessite des volumes de données très importants et que de telles quantités de données ne sont pas toujours disponibles pour les domaines spécialisés (domaines parfois très restreints ou bien dont les documents sont confidentiels) (Boleda 2020).

Pour cette raison, et d'autres, les outils de TAL basés sur l'apprentissage profond peuvent avoir des difficultés à prendre en compte la variation dans les domaines spécialisés.

2. Étude de cas

Cette étude va nous permettre de rendre compte d'un phénomène de variation concernant un fonctionnement lexico-syntaxique dans un domaine spécialisé, celui de la pêche. Nous verrons ensuite dans quelle mesure une étude de ce type pourrait être assistée par des méthodes d'IA.

2.1 La problématique

Nous abordons ici la présentation d'une étude dans laquelle Internet a été utilisé comme corpus, dans une approche en partie outillée.

Il s'agit de l'étude de la variation de la construction entre *pêcher* et *rivière(s)*, avec ou sans préposition. C'est par hasard que nous avons rencontré la construction sans préposition (*j'ai déjà pêché cette rivière*). Cet énoncé nous a paru échapper à la norme qui voudrait que *pêcher* soit suivi d'une préposition (*en, dans, sur*) pour introduire le complément de lieu. D'emblée, ce choix de construction nous a semblé propre aux pêcheurs et, par ailleurs, marquer une volonté inconsciente de rendre compte d'un lien privilégié avec la rivière, ce qui justifiait la suppression de la préposition. Cette hypothèse d'une variation sémantique en lien avec une variation de forme fait écho à celle de la grammaire constructionnelle : « *[…] in construction-based theories, grammar consists of a structured inventory of pairings of form and meaning* » (Goldberg 1996 : 8).

2.2 Deux types d'analyses

Deux types d'analyse ont été menés en utilisant les données disponibles sur Internet (l'exploration a débuté au cours de l'année 2012).

Dans la première étude, nous avons recherché toutes les occurrences

correspondant à [pêcher + préposition + (déterminant) + rivière(s)] ou [pêcher + déterminant + rivière(s)]. Seules les formes à l'infinitif ont été recherchées pour limiter le nombre d'occurrences à traiter. Les déterminants pouvaient être un défini ou un indéfini : *le, la, les, un, une, des* et les prépositions *dans, en, sur*.

À chaque occurrence, nous avons associé un type de site Internet : « pêche et subjectivité » ou « autre ».

Cette analyse a aussi été réalisée pour l'anglais en recherchant les structures [*to fish* + (déterminant) *river(s)*] et [*to fish* + préposition + (déterminant) + *river(s)*] avec les prépositions *on*, *within*, *in* et les déterminants *a*, *the*.

Dans la seconde analyse, nous nous sommes focalisée sur l'anglais et nous avons constitué deux sous-corpus, l'un avec les phrases contenant la structure avec préposition, l'autre avec les phrases contenant la structure sans préposition. Pour mettre en œuvre une approche lexicométrique, nous avons utilisé le logiciel AntConc (Anthony 2014) pour comparer les *keywords* significativement plus présents dans chacun des sous-corpus. Nous avons ensuite catégorisé sémantiquement « à la main » l'ensemble de ces *keywords* afin d'essayer d'établir un environnement sémantique caractéristique de chacune des constructions (avec *vs* sans préposition).

2.3 Résultats

Le tableau 1 rend compte de la répartition des structures avec préposition *vs* sans préposition, en français et en anglais, en fonction de la nature des sites.

	Tous les sites		Sites de pêche avec subjectivité	
	Français	Anglais	Français	Anglais
Structures avec préposition	82,3 %	49,7 %	61,6 %	30,8 %
Structures sans préposition	17,7 %	50,3 %	38,2 %	69,2 %

Tableau 1 : Structures avec *vs* sans préposition dans tous les sites *vs* dans les seuls sites de pêche avec subjectivité, pour le français et l'anglais

Les résultats présentés dans ce tableau confortent notre hypothèse de départ.

Pour le français, dans le corpus global, les occurrences sans préposition ne sont pas rares (17,7 %), contrairement à ce qu'un locuteur du français

standard (c'est-à-dire non pêcheur) pourrait penser. On peut constater de surcroît que la nature du site joue un rôle dans l'apparition d'une ou de l'autre structure ; en effet, dans les sites de pêche avec une dimension subjective, la construction directe est utilisée dans 38,2 % des cas. Précisons qu'un même site peut utiliser les deux structures. La mise en œuvre du Chi2 a confirmé l'existence d'une corrélation entre nature du site et apparition de la structure avec *vs* sans préposition.

Pour l'anglais, on peut noter que la structure sans préposition est présente presque à égalité avec la structure avec préposition (50,3 % *vs* 49,7 %). Les deux structures sont d'ailleurs admises et enregistrées dans les dictionnaires anglais. Par ailleurs, tout comme pour le français, la nature du site est en lien avec le choix d'une ou de l'autre structure : les sites de pêche avec une dimension subjective favorisent la présence de la construction sans préposition.

Voici quelques exemples de phrases avec ces deux constructions, pour le français et pour l'anglais :

1) Très jolie rivière ! tu as du bol de pouvoir pêcher une rivière aussi sauvage.
2) Avant de pêcher dans la rivière Northwest, les pêcheurs à la ligne doivent se procurer un permis de pêche du saumon du parc national.
3) I love to fish rivers, every single one I have ever been on is different.
4) To fish in the River Fowey you will need an Environment Agency Licence, (except under 12's), these can be purchased at a Post Office or by Telephone.

Pour ce qui concerne la seconde étude (menée seulement sur l'anglais), les résultats sont aussi très intéressants. Le tableau 2 rend compte des résultats chiffrés obtenus avec AntConc (Anthony 2014) pour les deux sous-corpus anglais. Rappelons que l'un de ces sous-corpus est constitué des paragraphes contenant le verbe « pêcher » suivi d'une préposition, l'autre des paragraphes contenant le verbe « pêcher » construit directement avec *rivière(s)*.

Le nombre de mots, mais aussi le nombre de lemmes, sont proches dans les deux corpus, ce qui a facilité la comparaison statistique.

	Corpus sans préposition	Corpus avec préposition
Nombre de mots	41361	40365
Nombre de lemmes	2715	2440
Nombre de lemmes significatifs (avec keyness > 3.84)	322	319

Tableau 2 : Résultats quantitatifs pour l'étude comparée des deux sous-corpus anglais

Parmi les lemmes spécifiques à l'un ou l'autre corpus, nous avons pu identifier 10 catégories sémantiques, 5 pour chacun des corpus.

Voici des exemples de lemmes associés à chaque catégorie ; pour le corpus « sans préposition » :

- Mois ou saisons : *January, February, spring, summer,*
- Poissons : *trout, pike, grayling, walleye, bream*,
- États ou régions : *Normandy, Nevada, Alabama,*
- Vocabulaire positif : *inspiring, ideal, beautiful, clarity, peacefulness*,
- Accessoires de pêche : *tackle, accessories, wader, bait, nymphs, line, braids* ;

pour le corpus « avec préposition » :

- Vocabulaire légal : *permission, unlawful, license, permit, law*,
- Vocabulaire économique : *property, owner, landowner, leaseholder*,
- Vocabulaire relevant du danger : *chemical, lethal, polluted, danger, arsenic, decrease, threat*,
- Éléments naturels : *cormorant, flower, animal, reef, plant, crocodile*,
- Relations familiales : *grandchildren, husband, ancestors, family*.

Les extraits 5) et 6) sont deux nouveaux exemples illustrant les deux catégories d'usage :

5) Should you be lucky enough to have the opportunity <u>to fish rivers</u> or streams the best flies to start with would be wet flies and nymphs.
6) You need a permit <u>to fish in the river</u>, but I am unsure of the cost.

Certaines catégories peuvent paraître étonnantes. Ainsi la surreprésentation des éléments naturels et la surreprésentation des relations familiales dans le corpus « avec préposition » donc « général ». Ces résultats peuvent peut-être s'expliquer par le fait que les pêcheurs sont surtout focalisés sur la rivière comme élément naturel et qu'ils préfèrent pêcher seuls (ou, peut-être, avec des amis). En revanche, ils sont particulièrement sensibles à leur environnement situationnel : le lieu et le moment où ils pêchent, mais aussi leur équipement et les types de poissons qu'ils peuvent rencontrer.

Ces deux études, rapidement présentées (pour plus de détails, voir Condamines 2017, 2018a, 2018b, 2021) mettent en évidence deux aspects.

D'une part, la construction sans préposition est plus utilisée par les pêcheurs que par les non-pêcheurs. D'autre part, lorsque cette construction

est choisie, elle s'accompagne d'un vocabulaire faisant intervenir la dimension affective ou émotionnelle. En témoigne en particulier l'utilisation d'un vocabulaire relevant du bien-être. Allant dans le sens de cette implication émotionnelle, on peut noter des énoncés dans lesquels la rivière est personnalisée, au point de faire l'objet de sentiments.

7) On a whim, I decided to fish a river in Oregon I had visited before, and fell in love with.

D'un point de vue méthodologique, notons que la recherche du lexique le plus significatif dans l'environnement des deux types de structures relève bien d'une analyse distributionnelle, focalisée sur le lexique présent dans l'entourage de chacune des structures, mais elle ne tient pas compte du rôle syntaxique de ces éléments.

Au-delà de la description linguistique, la description de ce phénomène a une importance pour comprendre le fonctionnement cognitivo-syntaxique des experts, qui n'est pas le fruit du hasard. Il faut donc aussi voir si ce phénomène de « transitivation du complément de lieu » se retrouve pour d'autres verbes et pour d'autres langues.

2.4 D'autres verbes, d'autres langues

2.4.1 D'autres verbes

En cherchant sur le web un peu au hasard, nous avons rencontré d'autres verbes admettant (même si c'est parfois très rare) la construction directe du complément de lieu, par exemple :

- Pagayer la rivière
  8) Tout sourire et solidement complices, Manon et moi allions enfin pagayer la rivière Kanasuta.
- Plonger la rivière
  9) Le chevalier Percevan va plonger la rivière souterraine qui alimente le puits où il était esclave.
- Skier la montagne
  10) Profiter de la poudre pour skier la montagne de Lure est devenu un évènement rare.
- Chasser la lande
  11) Henri VIII, qui aimait chasser la lande entourant le village, a donné le manoir à Anne Boleyn pour la vie.

Le cas de « *chasser* + nom de lieu » est particulièrement étudié dans Condamines (2018b).

En revanche, nous n'avons trouvé aucune occurrence de *tuer la rivière,* ni de *harponner la rivière* (avec des verbes qui peuvent se trouver à la place de *pêcher* devant *poisson*) ni de *danser une salle*, *boxer un ring*, *chanter La Scala/ L'Olympia*.

Cette différence entre les verbes pourrait être liée à la nature du lieu : *montagne, rivière, lande* renvoient à des lieux ouverts contrairement à *ring, salle, Olympia*. Mais il faudrait des études plus poussées pour vérifier s'il s'agit d'un aspect décisif dans l'impossibilité de la construction.

Il se pourrait aussi que l'origine territoriale du locuteur joue un rôle. Pour le français, les Québécois pourraient, par exemple, utiliser plus facilement la construction sans préposition (l'exemple 9 est ainsi issu d'un site québécois). Plus généralement, Callies (2018) a noté, en anglais américain, une tendance à supprimer les prépositions par exemple dans *he graduated Stanford* au lieu de *he graduated from Stanford*. Il faudrait envisager une catégorisation de l'origine des sites Internet pour pouvoir prendre en compte cette caractéristique, ce qui est loin d'être toujours faisable.

2.4.2 D'autres langues

Nous avons mené le même type d'étude, concernant *pêcher* et ses traductions, avec d'autres langues : l'espagnol, l'italien et l'occitan (en collaboration avec des locuteurs de ces langues). Le tableau 3 rend compte du nombre d'occurrences trouvées sur le web pour chaque structure dans différentes langues.

	Avec préposition	Sans préposition	Total
Anglais	1094	1108	2202
Espagnol	1226	384	1610
Français	1213	261	1474
Italien	961	26	987
Occitan	79	0	79

Tableau 3 : Présence *vs* absence de préposition dans les constructions dans différentes langues

12) Me cuesta abandonar sus orillas y mas el último día de temporada, sabiendo que no volveré a pescar el río hasta dentro de unos cuantos meses (espagnol).

13) I periodi migliori per pescare il fiume Snake sono all'inizio della primavera prima del ballottaggio (italien).

14) L'Ib-Salut esperarà esdeveniments i mai no jugarà la tàctica de pescar dins un riu remogut (occitan).

On peut noter que les deux constructions sont présentes pour toutes les langues hormis l'occitan. Pour cette langue, le nombre d'occurrences, même pour la construction canonique, est très faible (79). La question des langues régionales, faiblement présentes sur l'Internet, constitue un des problèmes majeurs des études linguistiques qui voudraient mettre en œuvre des méthodes d'apprentissage profond qui nécessitent des données volumineuses, nous l'avons déjà vu.

Dans une autre étude (Condamines 2017) nous avons pu voir que la nature du site (+ *pêche* + *subjectif*) jouait un rôle en espagnol comme en français et en anglais. En revanche, pour l'italien, langue pour laquelle la construction directe est rare, on ne retrouve pas cette corrélation.

2.5 L'IA pourrait-elle assister le travail des terminologues textuel(le)s ?

Dans cette dernière sous-partie, en prenant l'exemple de l'alternance de construction des compléments de lieu construits avec *vs* sans préposition, nous nous demandons quel pourrait être l'apport des systèmes d'IA basés sur des corpus volumineux. Du côté des apports possibles de l'IA, nous pouvons noter :

- La possibilité de rechercher les formes verbales non seulement à l'infinitif mais à tous les temps possibles, ce qui augmenterait le nombre de données utilisables. Il faudrait cependant pouvoir travailler sur un corpus étiqueté en parties du discours.
- La possibilité de faire porter la recherche sur d'autres verbes que le verbe « pêcher ».
- La possibilité de repérer automatiquement, pour un même verbe, les arguments construits directement *vs* construits indirectement.

La possibilité d'établir une corrélation entre nature de la construction et nature du site, avec un bémol concernant le fait qu'il peut être difficile d'établir précisément la nature du site (*i.e.* le genre textuel dont il relève). Ces deux derniers points, s'ils s'avéraient réalisables, constitueraient une réelle aide pour l'analyse linguistique.

Du côté des difficultés, nous pouvons noter :

- La nécessité de disposer d'un grand nombre de données ; or, nous l'avons vu pour l'occitan par exemple, certaines langues sont peu dotées du point de vue de la quantité de textes sur Internet.
- La difficulté pour les systèmes d'IA de repérer la nature sémantique des compléments qui sont construits avec ou sans préposition (par exemple le lieu dans le cas de *pêcher*). Des méthodes de sémantique distributionnelle en corpus général pourraient sans doute être convoquées pour contribuer à cette tâche.
- La nécessité, pour les linguistes-terminologues, de connaître les (très complexes car basés sur des savoirs mathématiques) systèmes d'IA et leur mode de programmation pour les adapter à une question nouvelle, ou bien de faire appel à des informaticiens spécialisés dans le TAL (ce qui les rend dépendants d'un tiers).
- Le fait que la question de la corrélation entre différents types de phénomènes (ici, la nature d'une construction et la nature du site) relève d'une intuition linguistique et pas d'une proposition de l'outil. De fait, le nombre des corrélations potentiellement repérées par des outils est très élevé et beaucoup n'ont pas de sens pour le linguiste-terminologue. Imaginons par exemple que l'outil découvre que les pêcheurs utilisent plus la lettre « e » que les non pêcheurs. À cette corrélation, le linguiste-terminologue ne peut donner aucune interprétation pertinente pour son point de vue. Et il n'y a sans doute aucune interprétation à donner.

En fait, on pourrait espérer qu'un outil contribue à répondre à la question : quels sont les verbes qui ont des constructions différentes (présence *vs* absence d'une préposition) pour les mêmes arguments et dans quelle mesure la nature du site joue-t-elle un rôle significatif dans cette alternative ? Cette question de la nature du site permettrait de prendre en compte des éléments extralinguistiques *via* la notion de genre textuel, voire des éléments relevant d'une cognition incarnée *via* l'étude du lexique. En outre, l'objet d'observation ainsi constitué se rapprocherait d'un corpus et donc d'un véritable objet d'étude linguistique et pas seulement de données car, comme le rappelle Rastier : « les données n'ont pas de sens » (Rastier 2021 : 232). Mais les outils ne pourraient sans doute pas proposer tout seuls d'aller regarder du côté de telle corrélation, qui semble présenter un intérêt linguistique. En d'autres termes, l'IA ne peut pas (en tout cas pour l'instant) remplacer l'intuition des linguistes, c'est-à-dire des experts de la langue qui élaborent des hypothèses.

Conclusion

Cet article a eu pour but de dresser un panorama des relations entre IA, terminologie et psychologie cognitive afin de voir ce qu'elles peuvent apporter à la réflexion sur la prise en compte de la variation dans les langues spécialisées. La présentation de l'étude sur la variation de construction (avec préposition *vs* sans préposition) entre *pêcher* et *rivière* et *to fish* et *river* a montré un lien entre la nature des sites Internet et la nature de la construction. L'article a aussi montré que la dimension émotionnelle des experts (les pêcheurs) pouvait être repérée par la significativité sémantique de l'environnement lexical des deux structures (présence importante du lexique du bien-être dans les phrases contenant les structures sans préposition). Cet aspect va dans le sens des propositions de la psychologie cognitive concernant la cognition incarnée, c'est-à-dire la prise en compte de la situation dans l'apprentissage et le fonctionnement du lexique mental. Nous avons aussi essayé de montrer comment les nouvelles méthodes d'IA pouvaient (ou ne pouvaient pas) contribuer à la mise au jour de ces fonctionnements.

L'évolution rapide des méthodes d'IA utilisées en TAL, désormais surtout basées sur la sémantique distributionnelle et l'apprentissage profond, a abouti à des résultats spectaculaires, par exemple dans la traduction automatique. Mais elle a aussi déstabilisé les experts de la langue, qui se sont sentis dépossédés de leurs connaissances et de leurs compétences puisque les outils fonctionnent sans connaissance linguistique. Passée cette étape de déstabilisation, les linguistes doivent réfléchir à la position qu'ils/elles peuvent avoir par rapport à cet état de fait. Au-delà de la seule production de ressources pour alimenter les systèmes de TAL, clairement en perte de vitesse, les linguistes pourraient prendre en compte les possibilités de ce nouveau TAL afin d'évaluer s'ils/ si elles peuvent, et à quelles conditions, intégrer ces méthodes dans leur objectif qui est, *in fine*, de comprendre les fonctionnements langagiers (essentiellement sémantiques) pour répondre, éventuellement, à des besoins sociétaux.

Bibliographie

Dernière consultation : 10 septembre 2021.

Anthony Lawrence (2014). AntConc (Version 3.4.3). Tokyo, Japan : Waseda University. http://www.laurenceanthony.net/software

Auger Alain, Barrière Caroline (éds) (2008). « Pattern Based Approaches to Semantic Relation Extraction: a State-of-the-art ». *Terminology*, 14/1, 1-19.

Aussenac-Gilles Nathalie, Condamines Anne (2007). « Corpus et terminologie ». In : Roger T. Pédauque (éd). *La redocumentarisation du monde*, Toulouse : Cepadues, 131-147.

Barsalou Lawrence (2003). « Situated Simulation in the Human Conceptual System ». *Language and Cognitive Processes*, 18, 513-562.

Baumann Klaus-Dieter (2007). « A Communicative-cognitive Approach to Emotion in LSP Communication ». In : Kurshid Ahmad, Margaret Rogers (éds). *Evidence-based LSP*. Berne : Peter Lang, 323-344.

Boleda Gemma (2020). « Distributional Semantics and Linguistic Theory ». *Annual Review of Linguistics*, 6/1, 213-234.

Callies Marcus (2018). « Patterns of Direct Tranzitivation and differences between British and American English ». In : Mark Kaunisto, Mikko Höglund *et alii* (éds). *Changing Structures, Studies in Constructions and Complementation*. Amsterdam/ Philadelphia : John Benjamins, 151-167.

Condamines Anne, Picton Aurélie (2022). « Textual Terminology: Origins, Principles and New Challenges ». In : Marie-Claude L'Homme, Pamela Faber (eds). *Theoretical Approaches to Terminology*. Amsterdam/ Philadelphia : John Benjamins.

Condamines Anne (2021). « How Can One Explain "Deviant" Linguistic Functioning in Terminology? ». *Terminology,* first published on line https://doi.org/10.1075/term.20029.con

Condamines Anne (2018a). « Is "To Fish in a River" Equivalent to "To Fish a River"? A Study at the Crossroads of Cognitive Sociolinguistics and Corpus Linguistics ». *Cognitive Linguistic Studies*, 5/2, 208-229.

Condamines Anne (2018b). « La transitivation des compléments circonstanciels dans le sport et les loisirs, en situation d'implication affective : néologie sémantique ou simple variation argumentale ? ». In : Delphine Bernhard, Maryvonne Boisseau *et alii* (éds). *La néologie en contexte. Cultures, situations, textes*. Limoges : Lambert-Lucas, 217-230.

Condamines Anne (2018c). «Terminological Knowledge Bases ». In : Pedro Fuertes-Olivera : *The Routledge Handbook of lexicography*. London : Routledge, 335-349.

Condamines Anne (2017). « The Emotional Dimension in Terminological Variation : the Example of Transitivization of the Locative Complement in Fishing ». In : Patrick Drouin, Aline Francoeur *et alii* (éds). *Multiple Perspectives on Terminological Variation*. Amsterdam/ Philadelphia : John Benjamins, 11-30.

Condamines Anne (2002). « Corpus Analysis and Conceptual Relation Patterns ». *Terminology*, 8(1), 141-162.

Condamines Anne, Picton Aurélie (à paraître)**.** « Textual Terminology : Origins, Principles and New Challenges ». In : Marie-Claude L'Homme, Pamela Faber (éds). *Theoretical Approaches to Terminology.* Amsterdam/ Philadelphia : John Benjamins.

Faber Pamela (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin : Mouton de Gruyter.

Fabre Cécile, Hathout Nabil, Sajous Franck, Tanguy Ludovic (2014). « Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille ». In : Actes de la *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, juin 2014, Marseille, France. Marseille : Université d'Aix Marseille, 266-279.

Firth John Rupert (1957). *Papers in Linguistics 1934–1951*. Oxford : Oxford University Press.

Firth John Rupert (1968). « Linguistic Analysis as a Study of Meaning ». In : Frank Robert Palmer (éd). *Selected Papers of J.R. Firth*. Londres : Longman (première édition 1952), 12-26.

Gaudin François (2003). *Socioterminologie – Une approche sociolinguistique de la terminologie*. Bruxelles : De Boeck - Duculot.

Glenberg Arthur, Robertson David (2000). « Symbol Grounding and Meaning : A Comparison of High-Dimensional and Embodied Theories of Meaning ». *Journal of Memory and Language*, 43, 379-401.

Goldberg Adele (1996). « Jackendoff and Construction-based Grammar ». *Cognitive Linguistics*, 7/1, 3-19.

Gries Stefan (2015). « The Role of Quantitative Methods in Cognitive Linguistics : Corpus and Experimental Data on (relative) Frequency and Contingency of Words and Constructions ». In : Jocelyne Daems, Eline Zenner *et alii* (éds). *Change of paradigms - New paradoxes: Recontextualizing Language and Linguistics*. Berlin/ Boston : Walter de Gruyter, 311-325.

Harris Zellig (1954). « Distributional Structure ». *Word*, 10(23), 146-162.

Heylen Kris, Bertels Ann (2016). « Sémantique distributionnelle en linguistique de corpus ». *Langages*, 201/1, 51-64.

Kristiansen Gitte, Dirven René (éds) (2008). *Cognitive Sociolinguistics : Language Variation, Cultural Model, Social Systems*. Berlin : Mouton de Gruyter.

Landauer Thomas (1999). « Latent Semantic Analysis (LSA), a Disembodied Learning Machine, Acquires Human Word Meaning Vicariously from Language Alone ». *Behavioral and Brain Sciences*, 22/4, 624-625.

Lenci Alessandro (2008). « Distributional semantics in linguistics and cognitive research ». *Italian Journal of Linguistics*, 20/1, 1-31.

Marshman Elizabeth, L'Homme Marie-Claude, Surtees Victoria (2008). « Portability of cause-effect relation markers across specialized domains and text genres : A comparative evaluation ». *Corpora*, 3/2, 141-172.

Meyer Ingrid (2001). « Extracting Knowledge-Rich Contexts for Terminography : A Conceptual and Methodological Framework ». In : Didier Bourigault, Marie-Claude L'Homme *et alii* (éds). *Recent Advances in Computational Terminology*. Amsterdam/ New York : John Benjamins Publishing Company, 279-302.

Meyer Ingrid, Bowker Lynne, Eck Karen (1992). « Cogniterm: An Experiment in Building a Terminological Knowledge Base ». In : *Proceedings of 5th EURALEX International Congress on Lexicography*, Tampere, Finland. Tampere : Studia Translatologica, 159-172.

Rastier François (2021). « Data vs corpora ». In : Damon Mayaffre, Laurent Vanni (éds). *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*. Paris : Honoré Champion, 203-246.

Segui Juan (2015). « Évolution du concept de lexique mental ». *Revue de neuropsychologie*, 7/1, 21-26.

Temmerman Rita (2000). *Towards New Ways of Terminological Description. The Sociocognitive Approach*. Amsterdam/ Philadelphia : John Benjamins.

Turney Peter David, Pantel Patrick (2010). « From Frequency to Meaning : Vector Space Models of Semantics ». *Journal of Artificial Intelligence Research*, 37, 141-188.

Wüster Eugen (1976). « La théorie générale de la terminologie - un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et les sciences des objets (Trans.) ». In : Henriette Dupuis (éd), *Essai de définition de la terminologie. Actes du colloque international de terminologie*. Québec : Régie de la Langue Française, 49-57.

De Europa Première partie - Special Issue 2022

Human-machine interaction: how to integrate plain language rules in the revision cycles of Neural Machine Translation output

Christopher Gledhill, Maria Zimina

Introduction

In this paper we explore one major impact of recent advances in Neural Machine Translation (NMT) on a Master's degree-level course in translation at Université de Paris: the need to teach advanced revision skills on the basis of NMT output. We argue that a key component of these skills is the ability to address the quality requirements of end-users relating to normalised language, in particular Plain Language (PL). PL is now required by the communication policies and Style Guides of many organisations. Yet the principles of PL are not always well defined, nor implemented uniformly in source texts. As a consequence, NMT output can be deficient in terms of PL, necessitating many cycles of post-editing and revision. For trainee translators, who are also learning to revise and edit texts in their non-primary language, getting to grips with the principles of PL can be an added difficulty.

In the following discussion, we share our experience of integrating the principles of PL into the teaching workflow of a website translation project (a second year Masters course called "*Traduction de site web vers l'anglais*" delivered in French, hereafter TSA). We start off by highlighting the relevance of PL from our perspective as practitioners and teachers of specialised translation. We then discuss the problems of integrating PL and other competencies into a translation course which uses a NMT platform as its methodological focus. This leads us to examine how the traditional workflow of post-editing/ revision cycles has been profoundly affected by the affordances offered by NMT in terms of customisation and translation memory matching. Finally, we conclude by weighing up the

Christopher Gledhill, Université de Paris 7, christopher.gledhill@u-paris.fr
Maria Zimina, Université de Paris 7, maria.zimina-poirot@u-paris.fr

challenges of building PL into both the teaching and NMT workflow. We suggest here that PL is an overlooked feature in current debates on the impact of NMT, and that educators, project managers and trainee translators need to treat the integration of PL as a challenge, but also an opportunity for adding value to the translation project as a whole.

1. Why Plain Language?

Plain Language originated as a rights campaign, whose advocates (such as the Plain Language Foundation[1]) have long argued for the principles of PL to be applied to official texts in administration, the law, medicine, etc., either from the point of view of citizens' or consumers' rights (campaigning for equal access to health care, readable contracts, etc.). It is significant that as a result of such lobbying, PL has been adopted as an editorial policy in many jurisdictions in the English-speaking world, for example the 2010 'Obama Law' in the USA, affecting federal and state communication with the public; the adoption of PL in all government communications in New Zealand (Cutts 2013), etc. Other language areas have also adopted the principles of PL, notably at the European level (as we see below) or at the national level (texts produced by the *Direction de l'information légale et administrative*, Benoît-Barnet *et alii* 2002).

We acknowledge that research on simplified language has never conclusively found that plain varieties of language are unequivocally beneficial for the producer of the text or the consumer, even though there are undeniable benefits in terms of document harmonisation (O'Brien 2003). Having said this, and given the widespread influence of PL in administration, the law and other areas, it appears to us that language specialists should at least be aware of the concept of PL, and if possible, should acquire knowledge of how to write clearly as a valuable transferable skill. This has certainly been the argument made by legal rewriters such as Balmford (2002) and the many commercial organisations offering to redraft technical texts on demand, as well as offering certification in PL for technical communicators[2].

It is perhaps worth adding that Plain Language is not Controlled Language (CL). CLs are designed for a specific form of technical communication (including spoken and written varieties) and are prescribed in expli-

[1] Plain Language Foundation: https://www.plainenglishfoundation.com/ (accessed 9 July 2021).
[2] In particular, Balmford was the founder of <cleardocs.com>.

cit linguistic terms (O'Brien 2003; Hartig, Lu 2014), PL on the other hand has no linguistic definition. We have previously attempted to set out some of the features of PL summaries written for non-expert medics[3]. On the basis of corpus data, we found clear evidence for a distinctive plain style, which contrasts with the style of abstracts written for experts. However, this descriptive approach is not widespread: for the most part, when editorial guidelines discuss PL, their recommendations are usually limited to a few general principles (Jelicic *et alii* 2016). Since PL is not a formally defined variety of language, but rather a discourse practice or a set of communicative goals, we can only propose a functional definition for the term: a conscious stylistic strategy designed to present expert knowledge to non-experts in clear style, that is to say in an accessible, user-oriented and self-contained form.

A lack of formal definition should not be seen as an obstacle: as we demonstrate below, there are good practical reasons why trainee translators and editors cannot and should not be tied to one language variety, largely because they are dealing with many different types of text, and thus require a greater degree of flexibility than that afforded by a strictly defined CL.

2. Style Guides and the requirement to adopt Plain Language

We start our discussion by examining the concept of the 'Institutional Style Guide', looking specifically at how these texts attempt to normalise language output, especially in terms of Plain Language. Of the 35 competencies that the European Master's in Translation (EMT) network defines as necessary for translation students following courses associated with the EMT network, several are relevant to the TSA course:

4 […] mastering […] presentation standards, terminology and phraseology […]
10 […] using the appropriate metalanguage and applying appropriate theoretical approaches
11 […] review and/or revise their own work and that of others
18 Master the basics of MT and its impact on the translation process […]
19 Assess the relevance of MT systems in a translation workflow […]
23 Work in a team, […] using current communication technologies […]
29 Clarify the requirements, objectives and purposes of the client […]

[3] That is to say the linguistic characteristics of Plain Language Summaries of expert medical texts (Systematic Reviews) written for non-expert decision-makers (Gledhill *et alii* 2019).

In this paper, we are interested in skills such as 18 and 19 above, but especially skill 5:

5 Implement the instructions, style guides, or conventions relevant to a particular translation

In other words, all professional translators should at least be aware of the concept of a Style Guide, and project managers should consider building the concept into their design of the project workflow. We also suggest to our students that if their future clients do not have their own Style Guide (some of whom work for small organisations, and are thus not aware of issues such as document or language normalisation), then it is the responsibility of the translator to point this out and propose a default guide of their own.

Assuming we need to integrate a Style Guide into our website translation course, the next question is: which one? There exist a number of freely-available guides which can serve as reference documents (e.g. the 80-page UNESCO Style Manual). However, they are often highly specific to the organisations which publish them. For the website translation project, we find it useful to refer to two sets of documents, which set out recommendations given to EU translators working in English:

1. The English Style Guide (ESG), produced by the Directorate-General for Translation, Brussels[4];
2. The Europa Web Guide (EWG), produced by the European Commission[5].

The English Style Guide (ESG) is relatively detailed, although only the first 60 pages (of 128 pages) are relevant to website translation, including topics such as Punctuation, Spelling, Capitalisation, Names, Titles, Numbers, Abbreviations, Units, Lists, Inclusive Language, etc. The Europa Web Guide (EWG) covers similar ground, but is shorter (25 subsections, each only one or two paragraphs long). Because our students are primarily involved in translating the websites of academic institutions and charitable associations, we give priority to the EWG over the ESG when there are differences in the recommendations (details such as list format, use of dashes, etc.).

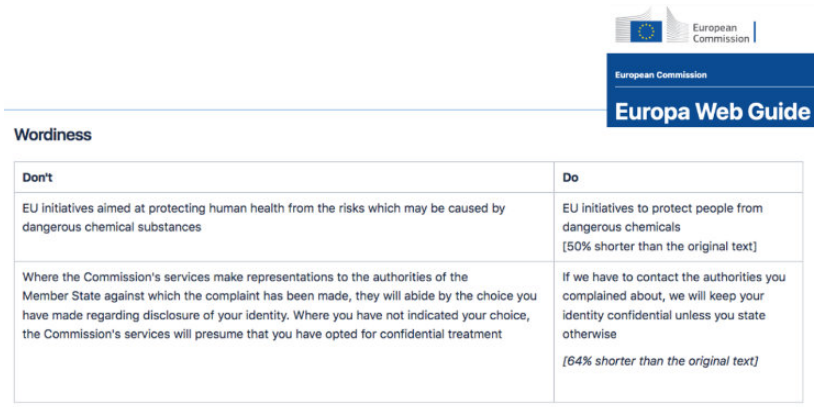
[4] English Style Guide: A handbook for authors and translators in the European Commission, https://ec.europa.eu/info/sites/default/files/styleguide_english_dgt_en.pdf (accessed 9 July 2021).

[5] Europa Web Guide: The official rulebook for the European Commission's web presence, https://wikis.ec.europa.eu/display/WEBGUIDE/02.+Style+guide (accessed 9 July 2021).

In all of the Institutional Style Guides that we have consulted, clear style is one of the first topics they deal with. The European Union has particularly embraced the concept of PL, producing a number of public relations documents about clear writing in various languages[6]. PL is consequently promoted as the preferred writing style in both the ESG and EWG. The EWG presents PL in highly positive terms, offering professional and ethical reasons for adopting PL, and reassuring the writer/ translator that PL is not going to make their text sound less elegant. One section of the EWG raises the issue of users with reading difficulties (a question that is addressed by the concept of Easy Language (Hansen-Schirra, Maaß 2020)). Another section "Jargon and plain language alternatives" presents a list of approved and disapproved terms. As far as linguistic structures are concerned, both ESG and EWG avoid detailed discussion (referring the reader to style manuals such as Cutts (2013)), although EWG presents some practical examples of reformulation. As can be seen in figure 1, this involves features that are often found in other types of Simplified English[7], such as:

- using active clauses instead of passives (*initiatives aimed at* > *initiatives to*),
- referring to animate participants instead of abstract ones (*initiatives, services* > *you, we, people*),
- using full verbs instead of light verbs (*make representation to* > *contact the authority*), etc.

As we can see in figure 1, the formatting of the example implies that it is possible to improve the target text by boiling it down to its essential


European Commission

European Commission

Europa Web Guide

Wordiness

Don't	Do
EU initiatives aimed at protecting human health from the risks which may be caused by dangerous chemical substances	EU initiatives to protect people from dangerous chemicals [50% shorter than the original text]
Where the Commission's services make representations to the authorities of the Member State against which the complaint has been made, they will abide by the choice you have made regarding disclosure of your identity. Where you have not indicated your choice, the Commission's services will presume that you have opted for confidential treatment	If we have to contact the authorities you complained about, we will keep your identity confidential unless you state otherwise *[64% shorter than the original text]*

Figure 1: Recommendation on 'Wordiness' in the Europa Web Guide

[6] For example, "Claire's Clear Writing Tips": https://ec.europa.eu/info/sites/default/files/clear_writing_tips_en.pdf, "*Rédiger clairement*" https://ec.europa.eu/info/sites/default/files/clear_writing_tips_fr.pdf (accessed 9 July 2021).

[7] For example, see the rules of ASD Simplified Technical English (ASD-STE100): http://www.asd-ste100.org (accessed 9 July 2021).

components. The informal metalanguage is also worth noting (*Lexical Density* > *Wordiness*). For these and other reasons, our students appear to find the EWG a very useful resource.

3. Integrating PL into the website translation project

We argued above that there is a clear need to raise awareness of PL among our trainee translators. However, this position raises a number of theoretical and practical questions which we would like to explore in the remaining sections of this paper.

Firstly, how can we integrate PL into our website translation course? As we see in the following sections, this is a course which requires students to acquire a broad range of specialised skills, including professional, technological and linguistic competencies. Obviously, it is not enough to simply present the recommendations on plain writing (as on the EWG webpage) and then move on to some other topic. Instead, we argue that it is necessary to actually build the principles of PL into the core structure of the teaching and assessment workflow.

However, any attempt to integrate PL into the teaching workflow raises a second question: how can we build PL guidelines into a workflow in which NMT is the central pillar? As we see below, there are many linguistic and textual features that are mentioned in PL guidelines that are difficult to be built into a NMT system, given the current state of the technology and the current methods by which NMT is implemented.

Finally, a related question concerns timing: at what point in our teaching workflow should we apply PL guidelines? During the whole project? Before the project begins? Or perhaps — an experiment we are currently trying out — during a final cycle of revision at the very end of the project, thus presenting PL guidelines as a final exercise of 'normalisation' or 'harmonisation'.

3.1 The webpage as text type: a challenge for our students and NMT

All text types vary in style, and few translators or linguists will be surprised to learn that a webpage mobilises very different discourse strategies and patterns of language from other types of text. We can illustrate this by taking an example from the website of the French Federation of Diabetics (FFD), one of our translation projects in 2020-2021.

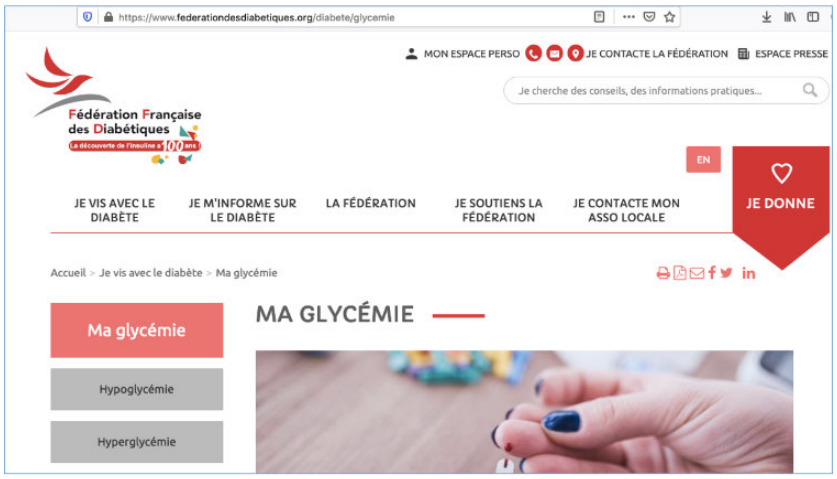


Figure 2: Source webpage in French (www.federationdesdiabetiques.org)

In figure 2 we can see that each lexical and grammatical choice in French is adapted to the interpersonal context of the website (calling for help, addressing the reader as a participant) as well as the textual context (fitting short phrases into small menu boxes). It is also notable that most of the information on this page has an intertextual function (helping the reader navigate somewhere else).

Thus, website translation presents many challenges including many varieties of discourse, and the average website project might involve translating the following:

- Menu items/main text items
- Links to interconnected websites
- Search forms, links to administrative documents
- Job adverts, lists of personnel, personal weblogs (blogs), biographies, obituaries, etc.
- Meta-texts, including guidelines, legal notices, contact information, etc.

It comes as a surprise to some of our students (especially those who have followed traditional translation courses), that they have to deal with all aspects of a webpage, including the *péritexte* (*via* text-navigation features). However, it soon becomes clear that there are just as many interesting translation problems to be solved, especially when it comes to post-editing and revising: the results of automatic translation of short text segments, such as menu items, demonstrate that the use of extended context is crucial to improve NMT technology (Tiedemann, Scherrer 2017).

In figure 3 you can see the FFD webpage (cf. Fig. 2) automatically translated from French into English using a generic neural MT engine: SYSTRAN Pure Neural® Server (SPNS)[8]. SPNS is a translation platform which seamlessly integrates with existing website architecture and helps our trainee translators handle website content translation/ revision without extensive knowledge of HTML/ XML formatting rules. It is deployed on a university intranet available to our Masters students. It is important to point out that this page was initially translated using a generic model (FR-EN) which had not been trained for a specific domain or topic. Thus a brief look at figure 3 reveals several examples of low-quality machine translation into English. For example, the informal abbreviation in French *"ASSO"* is inappropriately translated as "ASSO" in English. Another typical error concerns the menu items expressed as first-person clauses in French *("JE VIS"*, *"JE M'INFORME"*, *"JE CONTACTE"*, etc.). In particular, one of the main menu items on the French page of the association "[heart symbol] *JE DONNE*" (cf. Fig. 2) is clumsily translated by "[heart symbol] I GIVE" (cf. Fig. 3). But rather than point out such errors immediately, our question to the students is as follows: "Are these the best translations for these segments?". If not, we will clearly need to reconsider the other translation choices on the same page.

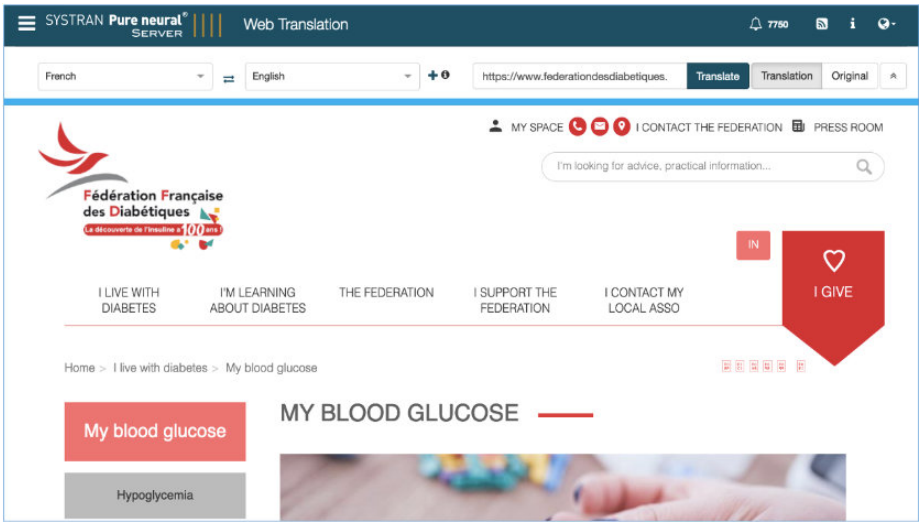


Figure 3: Initial NMT output in English (www.federationdesdiabetiques.org)

[8] See the official product website: https://www.systransoft.com/translation-products/systran-pure-neural-server (accessed 9 July 2021).

In the specific case of "[heart symbol] I GIVE", the students found that this is a typical discourse pattern on French charitable web pages, but not in English. Figure 4 shows how the page looks after our students researched alternative possibilities, including "DONATE" (an unfamiliar form of the verb) and "MAKE A DONATION" (which would be too long for the menu item, etc.) before they arrived at a participle verb ("DONATING"). This solution led to an interesting debate in the group about the tendency for English to prefer imperatives such as "SUPPORT" and "CONTACT" for links to procedural pages, as opposed to participles such as "LIVING WITH" for links to descriptive pages. Finally, it is notable that there is still a persistent error on the published page "I CONTACT THE FEDERATION" which requires further revision (a point which we return to below, cf. Fig. 10).

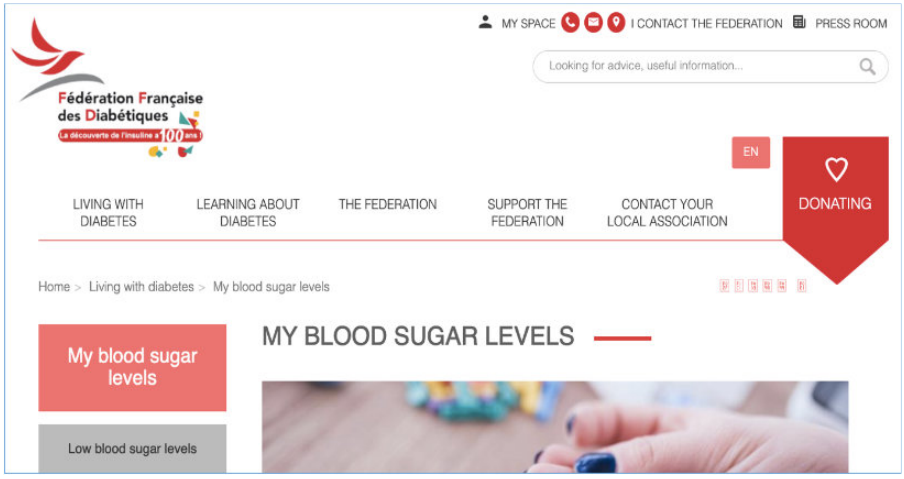


Figure 4: First revision of the webpage translation (www.federationdesdiabetiques.org)

Overall then, webpage translation is complex and requires several cycles of revision. Because every segment is context-sensitive, it follows that the quality and accuracy of NMT can vary significantly from one text segment to another, especially if the model used is untrained for this type of text.

It should come as no surprise therefore that, given such problems, as well as the sheer volumes of text involved, not all website owners are able to maintain high quality standards of translation, especially if they have to update their pages regularly, or deal with multiple languages[9]. In this

[9] We would like to thank an anonymous reviewer who mentioned the relevance of single-sourcing software (Hysell 2001) for website translation projects. Although this resource would be clearly relevant to this kind of project, not all institutional websites rely upon this type of software design.

case, an exclusion or limitation clause, or disclosure strategies are used to indicate that the MT output may contain errors of content or form. It is also considered good practice to clearly indicate that this translation process involves no human post-editing or revision. This type of statement can even be found on the European Union website regarding the use of eTranslation:

> Machine translation can give you a basic idea of the content in a language you understand. It is fully automated and involves no human intervention. The quality and accuracy of machine translation can vary significantly from one text to another and between different language pairs. The European Commission does not guarantee the *[sic]* accuracy and accepts no liability for possible errors[10].

However, this is not the paradigm we attempt to follow with our students. We inform them that it is our policy to adhere to the following quality statement (as set out by the European Commission Directorate-General for Translation):

> DGT's Quality Management Framework calls for our translations to be fit for their intended communicative purpose to satisfy the expressed or implied needs and expectations of our direct customers, our partners in the other EU institutions, the end-users, and any other stakeholders.

Since our objective is to achieve these quality requirements for all of the different texts we translate, we propose a solution that we call "Qualitative Translation/ Revision Workflow" (QTRW). We set out the details of this process in the following two sections.

3.2 Getting the best out of NMT: from generic output to customised NMT

The first stage in implementing QTRW involves a discussion of the different tools used in the project, most notably the professional translation platform SYSTRAN Pure Neural® Server. SPNS offers a broad range of features relevant to the website translation project, including[11]:

[10] Use of machine translation on Europa. Exclusion of liability: https://ec.europa.eu/info/use-machine-translation-europa-exclusion-liability_en (accessed 9 July 2021).

[11] Financed by a *"Projet Pédagogique Université Paris-Diderot"* (five-year contract: 2019-2023). Funding extension granted within the framework of the call for projects "Scientific Platforms and Equipment", Université de Paris, 2021: PAPTAN/DL4MT@UP (Deep Learning for Machine Translation at Université de Paris).

- Managing collaborative translation
- Identifying website updates
- Maintaining consistency
- Instantly showing how translations will look online
- Enabling Neural MT specialisation with custom resources
- Using Translation Memory and User Dictionary
- Producing output translation in HTML/XML format
- Helping with localisation

We now explore some of these features to explain to what extent they can be used to influence writing style and to improve translation quality. First, it is important to stress that it is not enough to use generic NMT models to obtain quality output: rather it is necessary to create custom NMT workflow (Senellart *et alii* 2003). However, this requires effort and skills, in particular a thorough understanding of the available data and linguistic resources (such as training corpora, translation memories, terminology, etc.). Thus, our students need to be involved at every stage of the process of NMT implementation, as we demonstrate below.

Figure 5 shows a custom NMT profile that delivers a reasonable translation quality improvement using a selection of available linguistic resources (normalisation dictionaries, bilingual terminology, translation memory combined with a translation model (FR-EN) selected in the SPNS 'Profiles' menu.

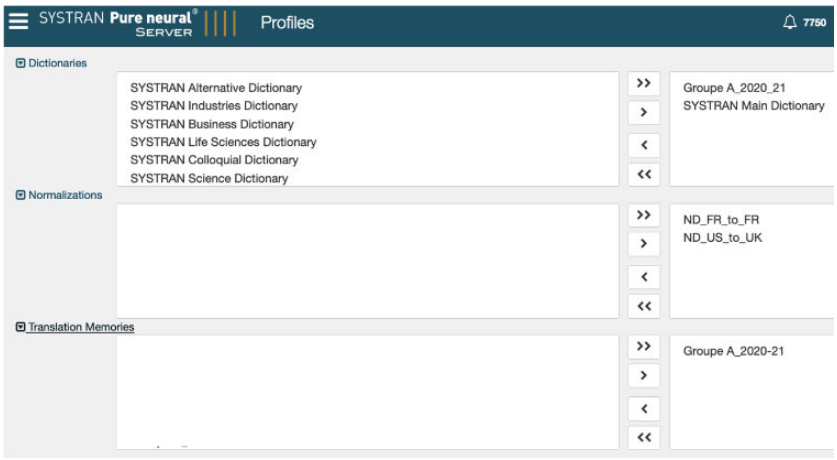


Figure 5: Neural MT specialisation: custom translation profile with selected linguistic resources

Along with rapid NMT customisation, a 'specialised' translation model can be developed to update an existing NMT engine, which will then deliver higher quality translation output than generic machine translation

(Servan *et alii* 2016). For example, figure 6 shows the output of generic NMT compared to that of 'specialised' NMT for medical discourse. It can be noted that the target segment "fasting and 2 hours after meals" becomes "on an empty stomach and 2 hours after meals" in the output of the domain-specific model.

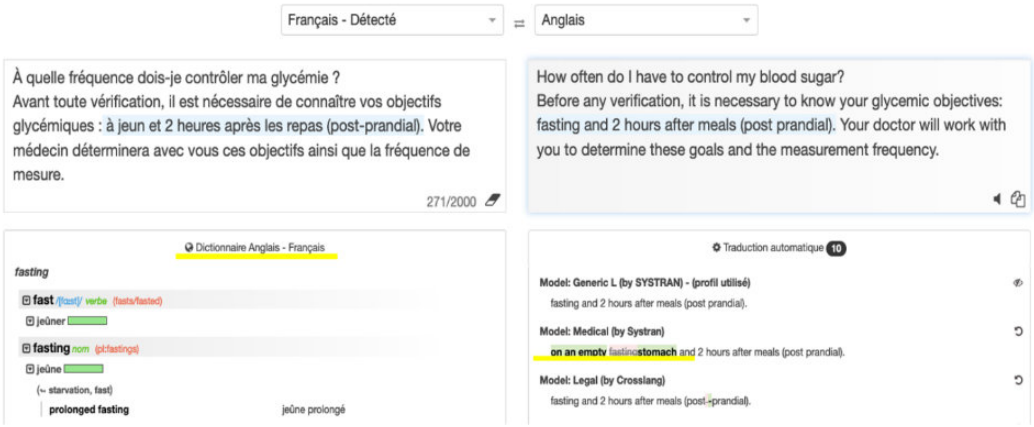


Figure 6: Generic NMT compared to that of 'specialised' NMT for medical discourse

As mentioned, customising NMT requires large amounts of domain-specific training data and resources. It is said that approximately 100 000 aligned text segments are needed to create a translation model from scratch[12]. Thus, every customisation is different, and requires a considerable amount of investment in terms of teaching time as well as learning curve.

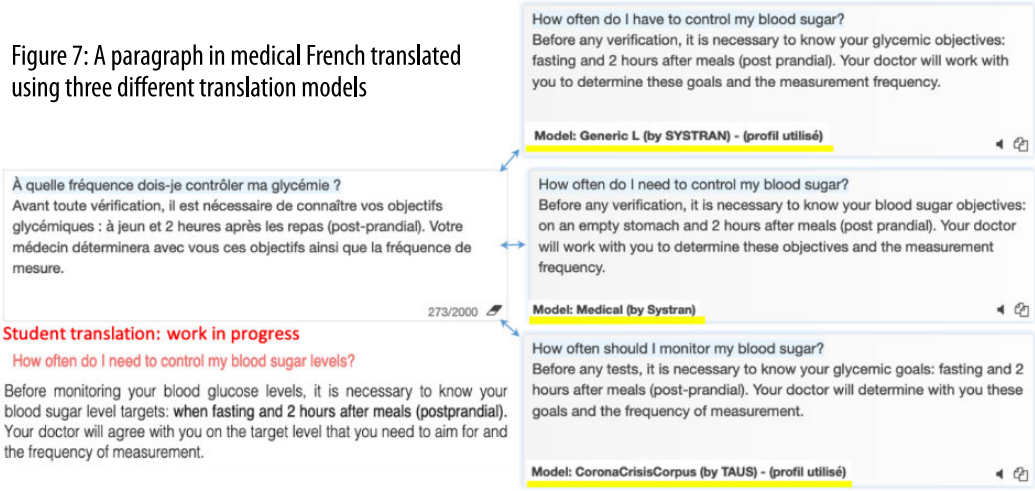


Figure 7: A paragraph in medical French translated using three different translation models

[12] Guidelines for successful corpus collection prior to Specialization. SYSTRAN Pure Neural® Server for End-Users Expert Users and Administrators. Systran, 24 February 2020, Université de Paris.

As further demonstration of NMT specialisation, figure 7 shows a paragraph from the FFD website translated into English using three different translation models. The first one is a generic model. The second was customised for medical English. The third is a domain-specialised model maximising access to factual information relating to the coronavirus. This example shows the impact of taking into account writing style. Figure 7 also shows that if the writing style does not match purpose, then the content must be revised (as you can see on the left-hand side in "Student translation: work in progress").

An additional point is that 'specialised' NMT models can be combined with normalisation dictionaries to correct spelling, units of measurement, formatting issues, etc. Such custom resources do not implement general linguistic patterns or rules. Normalisation allows for one-for-one replacements of the matched character strings in the source text (pre-editing step) and in the target text output (post-editing step). In this respect, state-of-the-art NMT is quite different from rule-based MT.

For example, figure 8 shows how UK spelling (a feature of normalisation that is imposed by current EU Style Guides) is implemented on a case-by-case basis.

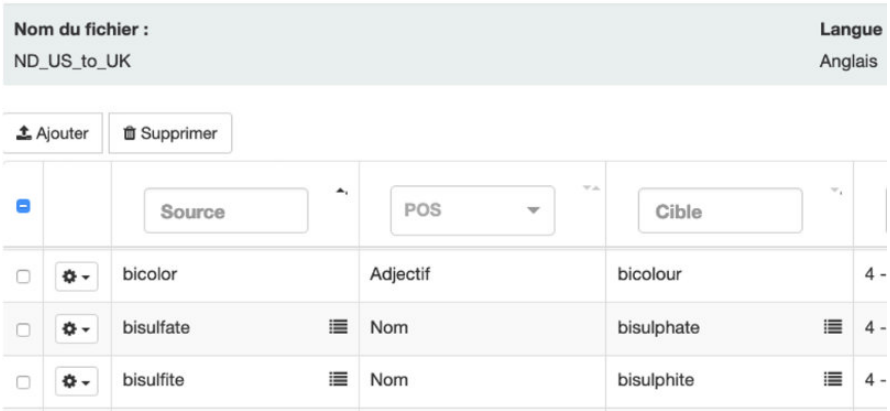


Figure 8: Custom NMT: SPNS Normalisation Dictionary for UK-EN spelling

Custom NMT models can also incorporate a User Dictionary (UD) to include project terminology and a bilingual glossary. The SPNS dictionary module implements one-for-one replacements with linguistic coding rules (singular, plural, verb conjugation, "do-not-translate"). Examples of this can be seen in figure 9, with items of project-specific terminology outlined in blue: the French-English UD entries are automatically propag-

ated from the bilingual glossary database to the translated text. The use of a fine-tuned customised model adhering to terminology constraints considerably reduces post-editing effort and improves translation consistency (Michon *et alii* 2020), especially in collaborative translation.

However, although they are particularly suitable for controlled languages, UDs are limited when it comes to polysemous terms with multiple translations.

Figure 9: Project-specific terminology in NMT output

Source	POS	Target
sommaire	Noun	contents
stylo	Noun	insulin pen
traitement oral/injectable	Expression	oral or injectable medication
Équilibre	Proper Noun	Équilibre

Pour mesurer votre glycémie, il vous faut :
• un lecteur de glycémie : l'appareil n'est remboursé par la Sécurité sociale tous les 4 ans qu'en cas de diabète traité par insuline ou de rétinopathie diabétique.
• des bandelettes ou électrodes (stockées à sec).
• un stylo autopiqueur à usage strictement personnel.
• des lancettes (fines aiguilles) : à usage unique, à adapter à l'autopiqueur.

392/10000

To measure your blood sugar levels, you need:
· a glucometer: the device is only reimbursed by the Social Security every 4 years in case of diabetes treated by insulin or diabetic retinopathy.
· strips or electrodes (stored dry).
· lancing device insulin pen for strictly personal use.
· lancets (thin needles): for single use, adapted to the lancing device.

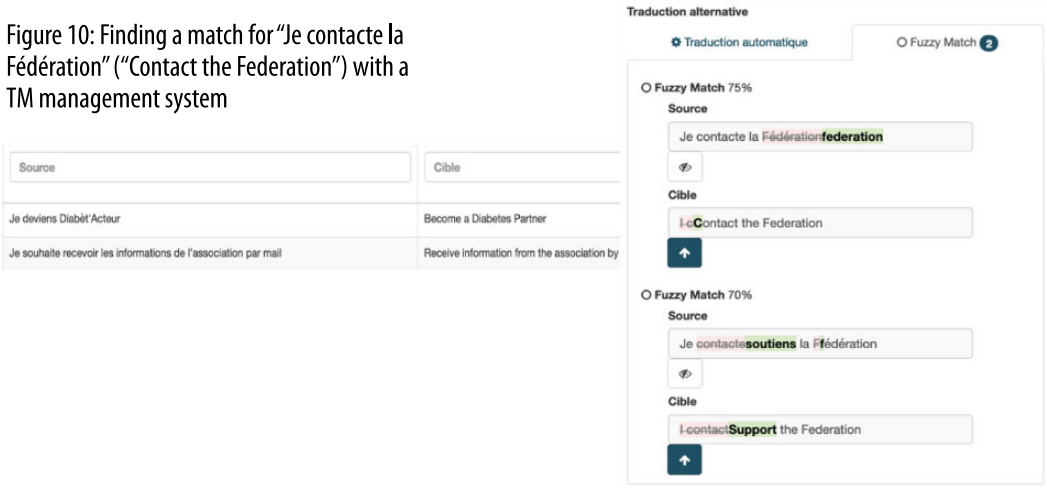
Finally, along with translation glossaries, it is possible to upload Translation Memories (TM) for a rapid customisation of a translation profile using database files (*.TMX) consisting of the original source texts and their respective translated segments. The machine translation system automatically reuses any segments from the database that are the same or similar (called 'fuzzy' matches) on new projects. TM resources are thus especially meaningful for human translators when it comes to NMT customisation. For group translation projects, sharing translation memory also guarantees total consistency of terminology, as it ensures that all translators are translating keywords and phrases uniformly[13]. However, before using a TM in a group project, an editor needs to review all included segments to ensure consistency.

In the future, new generation TM search engines may be able to identify similarity not only of lexical items but of longer stretches of discourse, such a routine lexico-grammatical patterns (Zimina 2019). For ex-

[13] On TM features in computer-aided translation: https://www.memsource.com/features/translation-memory (accessed 9 July 2021).

ample, figure 10 presents TM matches for the source segment identified in the revised version of the webpage that we analysed earlier ("*Je contacte la Fédération*" > "I contact the Federation", cf. Fig. 3).

Figure 10: Finding a match for "Je contacte la Fédération" ("Contact the Federation") with a TM management system



For segments to be reused, the TM management system relies on lexical similarities, in this case fuzzy matches of previously translated similar phrases, such as "*Je soutiens la fédération*". If the system encounters a novel segment, such as "*Je deviens Diabet'Acteur*" (literally, "I become a Diabetes Actor") or "*Je souhaite recevoir les informations*" (literally, "I want to receive information") which have both been translated using an imperative ("Become", "Receive", etc.), it should ideally be able to spot not just the formal similarity (no lexical matches in this case), but also the similar rhetorical function of these constructions.

The goal of the NMT search engine is to reduce human translator efforts by leveraging high-potential sentences through pattern matching. In other words, the translator should have the final say, but only when presented with a judicious selection of potential translation solutions. This point brings us back to the question of how to achieve efficient human-machine interaction within a Qualitative Translation/Revision Workflow.

3.3 Optimising human skills in the website translation project

It would not be accurate to divide a translation project like TSA into 'custom machine translation' on one hand and 'human intervention' on the other. As we have seen, our students are heavily involved in the project on both sides of the equation, and in two different but interdepend-

ent capacities. First, they have to be text analysts and data managers, in that they need to learn how to customise the NMT itself. Yet even after the custom NMT models have produced detailed and accurate output, there are still many tasks which are not 'solved' by the existence of NMT output, including revision, task management, delivery, and so on. Thus, our students have to be both language experts and project managers; they need to learn how to validate the quality of the NMT output, at the same time as complying with the specific quality requirements of the clients.

It is sometimes difficult for our students to appreciate both of these aspects simultaneously. One response to this problem has been to develop specific roles for each participant. Therefore, at the beginning of the project, we assign different job titles to each of the students, enabling the overall project to be broken down into manageable microtasks. For example, the auditing team determines the overall architecture of the website and decides which pages on the website can be translated; the terminologists manage not only technical terms but also the many repeated segments that need to be translated consistently across the website and implemented in the User Dictionary (such as the menu items we saw in paragraph 3.2); the task managers record which pages have been assigned to which translators and revisers, etc. In future iterations of the TSA course, we plan on asking the students to change roles, so that they will benefit from different vantage points on the project at different times.

Space precludes a full discussion of the many other aspects which need to be considered in terms of human resource management (as reported in Gledhill, Zimina 2019). So, we focus here on the one aspect of the project which requires considerable linguistic confidence on the part of the translators, and demonstrates very clearly the limits of machine translation, even intensively trained custom NMT: the need to re-write the target text in plain style.

As outlined above, we require our students to be familiar not only with the most relevant recommendations of two EU Style Guides (ESG and EWG), but also with the general principles of plain writing, which are implicit in these guidelines. We achieve this by asking our students to integrate many of the general principles of PL into the preparation of custom NMT (relevant terminology, spelling conventions, units of measurement, etc.). We also ask the students to act as editors and perform one 'final' cycle of revision (called 'language harmonisation') in which they are meant to apply the different recommendations of the Style Guide and PL

requirements to the latest version of the translated website. Once again, this task is broken down into microtasks, so that one team of students will be tasked with checking for gender-inclusive language (a requirement of the ESG), another team looks at excessive nominalisation (so-called 'zombie-nouns', as recommended by the EWG), etc.

There are two objectives to this exercise. Firstly, we want our students to adopt a systematic approach to the different types of linguistic errors and other types of recommendations that are involved in PL (and in the revision process in general, of course). Secondly, we want our students to be aware of the metalanguage in which these features are defined, and to adopt a systematic approach to classifying them. Although our students are often good at spotting inconsistencies, they are often excessively vague about the status of the different errors they have encountered, simply treating all problems as "not sounding right" (in English), *"bizarre"* (in French), etc. In an attempt to correct this 'blanket' approach to error analysis, we therefore ask them to categorise all of the edits they make in terms of four very broad categories, as set out here[14]:

a) Phonology-graphology
b) Lexico-grammar
c) Discourse-semantics
d) Context (of culture/ situation)

In the following subsections, we present examples of how different features of PL (or other items that belong to the requirements of our two style manuals: ESG and EWG) have been identified and classified by our students in previous translation projects (Gledhill, Zimina 2019). The point of this final exposition is to highlight the range of different categories of linguistic features that are simply not covered by NMT, thus requiring the linguistic skills of our students in identifying them as well as sometimes also finding appropriate translation solutions.

a) Revisions at the level of phonology-graphology[15]

This level of analysis covers problems of formatting, orthography, punctuation, typography, etc. Most of the recommendations in the EU

[14] The four 'strata' of language and other aspects of our metalanguage are based on Systemic Functional Grammar (Halliday, Matthiessen 2014).
[15] Although we are dealing with written documents, we retain Halliday's term 'phonology' so that we can potentially include audio translation. This also allows us to discuss the relation between punctuation and prosodic features, etc.

Style Guides are concerned with formal problems of this type, and most of the errors identified by our students are to be found at this level also. Superficially, it may seem that these features are trivial, but issues such as adopting the appropriate date format, respecting the conventions on capitalisation in English, the formatting of lists/ headings, etc. can collectively have an impact on text harmonisation.

Some issues are straightforward when viewed out of context (such as the requirement to write the date *14 April 2021* instead of *April 14th 2021,* etc.). But as the following example shows (PG1: taken from a student evaluation form), even the question of whether to use an initialism requires concentration and clear-thinking on behalf of the translator/editor:

> PG1
> Enriquillo-Plantain-Garden is used twice before the abbreviation EPG appears in brackets. From then on EPG is used. This is inconsistent. Changed English so that Enriquillo-Plaintain-Garden fault (EPGF) is used once, followed by its abbreviation. From then on, the abbreviation is employed. (J. W., TSA student project reporting form).

In addition, it is important for our trainee translators to be aware of the fact that 'high value texts' (such as the policy documents produced by the EU) must obey strict rules of formatting, especially when these texts are to be used as the primary texts for translation into many other languages.

b) Revisions at the level of lexico-grammar

As we go up one level, we arrive at the stratum of the lexico-grammar (the lexical and grammatical "resources for construing meanings as wordings" (Halliday, Matthiessen 2014: 131)). At this level many students are less sure of themselves in English, choosing to report simple morphological errors in the initial NMT output. However, more confident students are happy to engage with sophisticated issues of syntax and phraseology (selecting the appropriate combinations of words according to discourse type). Example LG1 shows a simple example where the student has chosen to apply the PL recommendations on 'avoiding the passive', as we can see in the following example (LG1, where FR: original French, EN0: machine translation output, EN1: first revision, etc.):

LG1
FR: *Le sud de l'île d'Haïti est traversée par une faille décrochante majeure : la faille d'Enriquillo-Plantain-Garden (EPGF).*
EN0: The south of the island of Haiti is crossed by a major stalling fault: Enriquillo-Plantain-Garden fault.
EN1: The Enriquillo-Plaintain-Garden fault (EPGF), a major lateral fault, cuts across southern Haiti [...]

Example LG2 shows a more sophisticated type of reformulation: 'de-nominalisation', the rewording of a process noun by a clause, thus making the participants in the clause explicit. LG2 is complex because it involves an increased degree of nominalisation in the NMT output (EN0 and EN1), followed by interventions by the translators/ editors who progressively reformulate the nouns as clauses (EN2). This example gives a good picture of how this type of sophisticated reformulation is typically introduced over a series of different cycles of revision rather than all at once:

LG2
FR: *Ces résultats amènent donc à repenser la structure et le fonctionnement de ces écosystèmes* [...]
EN0: These results lead to a rethinking of the structure and functioning of these particular ecosystems [...]
EN1: These results thus allow researchers to reconsider the composition and functioning of these specific ecosystems [...]
EN2: These results thus allow researchers to reconsider what these specific ecosystems were made of, and how they worked [...]

c) Revisions at the level of discourse-semantics

As we go up another level, we arrive at the level of discourse-semantics ("the set of strategies for construing, enacting and presenting non-language as meaning" (Halliday, Matthiessen 2014: 189)). This allows students to discuss problems of cohesion/ coherence as well as terminology (conceived here as exophoric reference). This level of analysis also includes reformulations above the level of the clause (seen as the upper limit for problems of lexico-grammar). The following example (DS1, identified by student M. H.) shows a distinct problem at this level of analysis: over-complex phrase structure, in this case involving multiple subordination in FR (which is carried over into EN):

DS1:
I found this sentence in French quite difficult to follow in order to understand the different steps. Therefore I decided to split the sentence so as to make the English version smoother. (M. H., TSA student project reporting form).

FR: *Toutes les lunes formées en dessous de l'orbite synchrone, dont Phobos, chutent vers Mars, mais Deimos s'étant formé juste au-dessus de l'orbite synchrone il est repoussé vers l'extérieur.*
EN0: All the moons formed below the synchronous orbit, including Phobos, fall to Mars, but Deimos having formed just above the synchronous orbit it is pushed back to the outside.
EN1: All moons that were formed below the synchronous orbit, including Phobos, are being pulled towards Mars. // Deimos, however, was formed just above the synchronous orbit and is therefore being pushed outward.
EN2: All the moons that were formed below the synchronous orbit, including Phobos, are being pulled towards Mars. // Deimos, however, was formed just above the synchronous orbit and is therefore being pushed outward.

As we can see in DS1, our students are usually very reactive when it comes to 'chopping' texts down to size like this (here signalled by a double slash **//**). And as discussed above, this type of intervention is entirely in keeping with the recommendations of the EWG on sentence length/complexity.

d) Revisions at the level of context

At the highest stratum of analysis, we have context of culture/ context of situation. These terms refer to errors which are significant because of cultural expectations, or due to a specific, transient feature of the situation[16]. Problems at this level are sometimes quite serious, but also consequently not frequent, and students sometimes have difficulty identifying them. With experience however, it is almost always possible to identify contextual problems, right across the website.

Example CS1 represents a very clear problem of cultural context, in that the NMT has assigned a gendered determiner "his" in English to translate the French possessive "*sa*":

CS1
FR: *une entreprise, qui confie à un doctorant un travail de recherche objet de sa thèse* ;

[16] Both concepts originate in the contextualist theory of J.R. Firth, as mentioned by Anne Condamines (this volume).

EN0: a company, which entrusts to a Ph.D. student a research work that will be the subject to his thesis;
EN1-2: a company which entrusts research work to a Ph.D. student on the subject of his/her thesis.

This gender-specific usage is proscribed in official documents in many English-speaking countries, and likewise the recommendations of both the ESG and EWG stipulate that when gender does not need to be specified, then the authors must adopt an alternative strategy (i.e. gender-neutral pronouns/determiners).

Example CS2 represents a different but also very typical problem, this time relating to the context of situation:

CS2
FR: *Les résultats seront affichés en ligne en janvier 2017*.
EN0: Results will be posted online in January 2017.
EN1-2: Results will be published online as soon as possible.

Here the source text has not been updated, and so the NMT output has just translated the out-of-date information as presented. In the absence of immediate feedback from the website owner, it is often necessary for our students to be proactive, and so in this case it was decided to provide a neutral 'holding' statement.

Finally, example CS3 shows a problem of editorial policy. As this relates to the expectations of the end-users and website owners, it is analysed as context of culture:

CS3
FR: *un récit des principales observations du Laboratoire de Physique du Globe effectuées* [...]
EN0: an account of the principal observations of the Physics laboratory of the Earth [...]
EN1-2: a narration of the main observations of the Laboratoire de Physique du Globe [...]

Generally speaking, generic NMT model will attempt to translate the names of institutions (especially less well-known ones). However, in this case it turns out that the organisation wants to maintain its identity and French name, so the translator/ reviser has to intervene. Here the NMT output can be adjusted with a User Dictionary, as explained in paragraph 3.2.

Conclusion

Neural Machine Translation (NMT) is a working tool, and is fast becoming the most important tool available in the translation industry. But the human translator still has a vital role to play. As we have set out in a number of instances in this paper, even when NMT output is customised, there can be significant discrepancies between this output and the principles set out in an Institutional Style Guide, in terms of cohesion, consistency, variations, standards and drafting codes, etc.

Controlled Languages are rule-based and can be implemented using machine translation fairly consistently. But as we have seen, the guidelines for Plain Language are essentially subjective and open to a degree of interpretation. It follows that if the customer's text is to be written in 'clear style', 'plain writing', etc., then the translators/revisers actively need to carry out a number of interventions. As demonstrated above, such edits are difficult to implement in NMT systems, and even if there were a means of customising the NMT to produce Plain Language output, this would in many ways defeat the object: custom NMT output is by definition adapted to a specific customer's purpose, while the concept of Plain Language is necessarily flexible, contingent and can vary not only from project to project, but even from one page of a website to another (and even conceivably from one section of the page to another...). For the time being, therefore, we must conclude that it is necessary to raise awareness of Plain Language among trainee translators: they need to learn how to mobilise both their knowledge of the language system and of editorial standards.

Regarding 'Human-Machine interaction', our present position is this: in any professional translation project such as the translation of a whole website, the place of the machine is to produce an initial translation. After this step, the role of revision as a process and as a human skillset is critical to the successful completion of such a project. But because of the increasing complexity of this process, it is just as important for trainee translators to acquire the appropriate skills in designing and implementing a translation/revision workflow that attempts at every turn to promote quality output, as it is for them to intervene pro-actively in project management and in the several complex cycles of post-editing and revision which are still necessary for this type of project. It is this dual relationship and responsibility which we have attempted to characterise here as a translation project design philosophy, and which we term "Qualitative Translation/Revision Workflow" (QTRW).

References

Last accessed 9 July 2021 (all the links in References).

Balmford Christopher (2002). "Plain language: beyond a movement". In: *Proceedings of the Plain Language Association International (PLAIN) Fourth Biennial Conference*, Toronto, Canada. September 26-29, 2002. Copian Library. Available online: http://en.copian.ca/library/research/plain2/movement/movement.pdf

Benoît-Barnet Marie-Paule, Collette Karine, Laporte Danielle, Pouëch Françoise, Rui-Souchon Blandine (2002). *Guide pratique de la rédaction administrative*. Paris: Ministère de la fonction publique et de la réforme de l'état.

Cutts Martin (2013). *Oxford Guide to Plain English*. Oxford: Oxford University Press.

Direction de l'Information Légale et Administrative (DGLFLF/OQLF) (2006). *Rédiger... simplement*. Québec: Editions de la Délégation générale à la langue française et aux langues de France/ Office québécois de la langue française.

Gledhill Christopher, Martikainen Hanna, Mestivier Alexandra, Zimina Maria (2019). "Towards a linguistic definition of 'Simplified Medical English': applying textometric analysis to Cochrane medical abstracts and their plain language versions". *Lingue Culture Mediazioni/ Languages Cultures Mediation*, 11, 91-114.

Gledhill Christopher, Zimina Maria (2019). "The Impact of Machine Translation on a Masters Course in Web Translation: From Disrupted Practice to a Qualitative Translation/Revision Workflow". In: *Proceedings Translating and the Computer 41*, November 2019. Editions Tradulex, Geneva, pp60-73. Available online: http://www.tradulex.com/varia/TC41-london2019.pdf.

Halliday Michael, Matthiessen Christian (2014). *Introduction to Functional Grammar*. 4th revised edition. London: Routledge.

Hansen-Schirra Silvia, Maaß Christiane (eds) (2020). *Easy Language Research: Text and User Perspectives*. Berlin: Frank and Timme.

Hartig Alissa, Xiaofei Lu (2014). "Plain English and legal writing: Comparing expert and novice writers". *English for Specific Purposes*, 33, 87-96.

Hysell Deborah A. (2001). "Single sourcing for translations". In: *Proceedings of the 19th annual international conference on Computer documentation (SIGDOC'01)*. Association for Computing Machinery, New York, NY, USA, 89–94. Available online: DOI:https://doi.org/10.1145/501516.501535

Jelicic Kadic Antonia, Fidahic Mahir, Vujcic Milan, Saric Frano, Propadalo Ivana, Marelja Ivana, Dosenovic Svjetlana, Puljak Livia (2016). "Cochrane Plain Language summaries are highly heterogeneous with low adherence to the standards". *BMC Medical Research Methodology*, 61/16. Available online: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0162-y

Michon Elise, Crego Josep, Senellart Jean (2020). "Integrating Domain Terminology into Neural Machine Translation". In: *Proceedings of the 28th International Conference on Computational Linguistics*, 3925–3937. Available online: https://doi.org/10.18653/v1/2020.coling-main.348

O'Brien Sharon (2003). "Controlling Controlled English. An Analysis of Several Controlled Language Rule Sets". In: *Proceedings of EAMT-CLAW-03*, Dublin City University: Dublin, 15-17 May 2003, 105-114.

Senellart Jean, Boitet Christian, Romary Laurent (2003). "XML Translation Workflow". In: *Machine Translation Summit IX*, September 2003, New Orleans, United States. Available online: https://hal.inria.fr/inria-00487747

Servan Christophe, Crego Josep, Senellart Jean (2016). "Domain specialization: A post-training domain adaptation for Neural Machine Translation". Submitted to EACL 2017. Available online: http://arxiv.org/abs/1612.06141

Tiedemann Jörg, Scherrer Yves (2017). "Neural Machine Translation with Extended Context". In: *Proceedings of the Third Workshop on Discourse in Machine Translation*, 82–92. Available online: https://doi.org/10.18653/v1/W17-4811

Zimina Maria (2019). "Vers une Mémoire de Traduction dynamique et multidimensionnelle". *Des mots aux actes*, 8, 221-236.

De Europa Première partie - Special Issue 2022

A Journey in Neural Machine Translation*

Philippe Langlais

Introduction

Machine Translation has dramatically improved over the last decade, mainly thanks to progress in neural machine translation (Sutskever *et alii* 2014; Bahdanau *et alii* 2015; Vaswani *et alii* 2017). It is now agreed that NMT is producing more fluent translations with much fewer errors than previous statistical machine translation (SMT) systems (Brown *et alii* 1993), owning to convincing comparative studies such as Isabelle *et alii* (2017).

Some studies have shown that the extra fluency of NMT does not necessarily warrant better post-edition. For instance, Castilho *et alii* (2017) show that post-editing the output of NMT in the educational domain did not save time compared to post-editing the output of SMT. Sánchez-Gijón *et alii* (2019) also reported empirical evidence that post-editing NMT was no faster than post-editing segments produced by a translation memory (TM) system.

On the contrary, Toral *et alii* (2018b) reported that post-editing the output of NMT conducted to doubling time savings, compared to post-editing SMT. Also, Läubli *et alii* (2019) compared the translation productivity of professional translators equipped with either a translation memory, or a mix of a TM for close matches and NMT otherwise. On the finance domain, they observe that NMT enabled professional translators to work faster for the German-French translation direction they tested (a speed gain of 59%), while being almost on par with the German-Italian language pair (a gain

Philippe Langlais, Université de Montréal, felipe@iro.umontreal.ca

* This paper expresses the views of the author only, but is informed by the work of several researchers we would like to thank. First, experiments have been conducted by amazing students it has been a pleasure to work with, namely Abbas Ghaddar, Shivendra Bhardwaj and Xavier Frenette. Second, some of the experiments described here have been performed in collaboration with colleagues at National Council of Canada, namely Michel Simard, Cyril Goutte and Gabriel Bernier-Colborne. We were also very lucky to have access to the translation memory in use at the Translation Bureau of Canada.

of 9% is measured). Furthermore, the authors found that NMT had no negative impact on the quality of the translations, on the contrary.

Obviously, more studies are required to have a better picture of the situation, notably checking the impact of more recent NMT engines, as well as neural systems that integrate a translation memory (e.g., Xu *et alii* 2020), as well as document-level ones (e.g., Lopes *et alii* 2020). Still, we believe that properly used, current NMT is a real asset in a professional setting. With this is mind, the purpose of our paper is to warn that maturing an NMT solution is not an easy endeavour. It requires a rather specific expertise as well as adequate computing facilities. Without this, an organization wishing to use NMT in its production pipeline has no other option than dealing directly with translation providers, that hopefully will take care of adapting the technology to the specificities of the needs.

This paper is organized as follow. In paragraph 1, we discuss the type of expertise required for training a neural engine, making use of dedicated sota libraries. We address the quality that neural engines produce in paragraph 2. In paragraph 3, we articulate that data is an important issue in developing a sound NMT solution. We discuss in paragraph 4 other options to the integration of neural translation that might prove more useful than simply post-edition, and conclude in paragraph 5.

1. Training an NMT system is no picnic

Progress in MT came with a number of packages[1] that are ready to use. We have been experimenting with three popular recent ones, that were all becoming prominent at the time we conducted the experiments.

XLM (Conneau, Lample 2019) is an architecture which tackles cross-lingual pre-training in a way similar to the BERT model (Devlin *et alii* 2019) with a few notable differences. First, XLM is based on a shared source-target vocabulary using Byte pair encoding (BPE) (Sennrich *et alii* 2016). Second, XLM is trained to predict both source and target masked words, leveraging both the surrounding words and the other language context, encouraging the model to align the source and target representations. Third, XLM stores the ID for the language and the tokens order (i.e., Positional Encoding) in both languages which leads to build a relationship between the related tokens in the two languages.

[1] https://awesomeopensource.com/projects/neural-machine-translation (last consulted: September, 28th 2021) lists 228 NMT packages.

We used the TLM architecture, and modified the original pre-processing code such that XLM can accept a parallel corpus for training TLM, which was not yet implemented in the GitHub at the time we conducted our experiments[2]. We used the 60k BPE vocabulary which comes with the pre-trained language model[3].

ConvS2S (Ott *et alii* 2019) is one predominant method for sequence-to-sequence learning which maps an input sequence to an output one of variable length *via* Recurrent Neural Networks (RNN/LSTM). The work of Gehring *et alii* (2017) demonstrates that Convolutional Neural Network (CNN) can be used for doing so. The ConvS2S model uses an encoder and a decoder that exploits a CNN with Gated Linear Units (Dauphin *et alii* 2017), non-linearity for both encoder and decoder and applies multi-step attention which is more elaborated than the attention typically used in RNN models. We used the implementation available in the fairseq toolkit.

We used a source and target vocabulary of 60K BPE types. The translation is generated by a beam-search decoder with log-likelihood scores normalized by sentence length.

Scaling-NMT (Ott *et alii* 2018) is a transformer model that showcased an improvement in training efficiency while maintaining state-of-the-art accuracy by lowering the precision of computations, increasing the batch size and enhancing the learning rate regimen. The architecture uses the big-transformer model with 6 blocks in encoder and decoder networks. The half-precision training reduced the training time by 65%. Scaling NMT is implemented in PyTorch and is also part of the fairseq toolkit[4].

We used the default 40K vocabulary with a shared source and target BPE factorization. During training and for translating, we use a beam search of width 4 and a length penalty of 0.6. For translation, we average the last five checkpoints.

From the description above, one can note that each package requires the setting of some meta-parameters. Their values can drastically impact performances, forcing the system developer to understand and explore packages to some level, a time-consuming exercise. As an illustration of this, it took almost 2 months of trial/ error iterations for an NLP engineer very familiar with deep learning to develop an NMT engine with satisfact-

[2] https://github.com/facebookresearch/XLM.git (last consulted: May, 3rd 2021).

[3] Training TLM without pre-training was rather unstable. We also noticed better results with a back-translation step, but at a high cost in training time.

[4] https://github.com/pytorch/fairseq (last consulted: May, 3rd 2021).

ory performance out of the packages considered. It required inspecting code, reading scientific papers, and even sometimes discussing with the authors of a package. Our point is not to criticize those packages (which are truly amazing), but to warn the novices[5] who believe that training an NMT engine is straightforward: packages are designed for computer scientists, often lack appropriate documentation and sometimes need some fixes. Also, since the technology is evolving fast, earlier packages may fail to deliver today's sota performance.

Another important point to note is that deep learning models definitely require a Graphical Processing Unit (GPU) to run, a facility which might not be available in all environments where the technology is to be used. Also, training can be quite time consuming. The average time to train XLM on four Tesla V100-SXM2 GPUs was around 22-30 hours depending on the training material and the number of epochs. For ConvS2S, it was 72-96 hours, and for Scaling-NMT, the training took 30-50 hours of computation on a single Titan V GPU.

2. Quality of neural machine translation

We are overwhelmed by media sources that praise neural machine translation (among other things). One of the earliest overly optimistic articles we read dates back to 2016 when NMT technology was just blooming and was published by *The Verge*[6] in a blog entitled "Google's AI translation system is approaching human-level accuracy" which is definitely misleading, as somehow acknowledged in the subtitle of this blog "But there's still significant work to be done" (Statt 2016). Two years later, many overenthusiastic blogs have been published when DeepL[7] and Quantmetry[8] gathered to produce the translation of the reference textbook on deep learning (Goodfellow *et alii* 2016), a document of 800 pages that was translated in no more than 12 hours[9]. Escribe (2019) inspected the raw output produced by the translation engine that was used

[5] It is one thing to run a package on a toy bitext, as demonstrated in many blogs, where a GPU is not even required. It is another kettle of fish to scale a system to larger datasets with state-of-the-art results.

[6] Many other such blogs have been published during this period, see for instance Zhou (2018).

[7] https://www.deepl.com/

[8] https://www.quantmetry.com/

[9] Not counting the time for developing the tool that could handle the markup language (LaTeX) the book was written in, but counting the time to post-edit the output of machine translation.

to do so, as well as the meta-data available on the post-edition conducted. She measured that 21% of the inspected segments where left unchanged, and noted that while most changes concerned terminology, many were actually only preferential. This is indeed an impressive achievement, since translating such a book would have required an entire year of a professional translator according to the author, at five times the cost.

The praise for neural MT is not specific to the blogosphere, and research studies are actually debating on the parity of neural MT with humans. Notably, Hassan *et alii* (2018) claimed NMT reached human parity for the translation into English of news in Chinese. However, in Toral *et alii* (2018a) the authors revisited this study, paying attention to overlooked variables (such as the proficiency of the evaluators) and found that parity was not achieved.

Again, the debate is not closed, and more studies are required to better appreciate the current state of affair. Our intent in the remainder of this section is admittedly naïve: we want to warn that current NMT engines are producing translations that differ from the human ones. Being different does not mean automatic translations are bad, but it remains to be studied if the differences are acceptable in a given context of use. Yet another task that should fall into any organization that seriously plans to make good use of NMT.

In Bhardwaj *et alii* (2021a), we have conducted experiments showing that it is possible to train a classifier to distinguish translations produced by a neural system from those made by professional translators, with high accuracy in some contexts. For instance, we could train deep learning classifiers to recognize neural translation produced by an XLM model we trained with an accuracy as high as 87.5% (random guessing would lead to 50%) when tested on material we gathered from the Translation Bureau of Canada. Admittedly, translation quality in this organization is very high, facilitating the work of the classifier. In fact, one of the most challenging settings we found was to discriminate translations produced by DeepL on sentences from the Canadian Parliament debates, where an accuracy of 66.5% was observed (still better than random guessing). It is notorious that DeepL has been trained on such material, therefore delivering translations of very good quality, especially since this type of text tends to be somehow repetitive.

SRC	HUM	NMT
a. TBBI evaluation scheduled for 2017-2018 to ensure FAA compliance (Hep C contributions).	L'évaluation des ITSS est prévue en 2017-2018, pour assurer la conformité à la LGFP (programme de contributions – hépatite C).	* Évaluation du STBBI prévue pour 2017-2018 pour assurer la conformité avec la FAA (contributions au titre de l'hépatite C).
b. 6c) Were you informed about the ADR process at the CHRC?	6c) Vous a-t-on informé du processus relatif au RAD de la CCDP ?	6c) Avez-vous été informé du processus de MARC à la CCDP ?
c. Are there any specific services being requested by SMEs that you are not able to provide for them or that you feel lie outside of your mandate?	Les PME vous demandent-elles de leur fournir des services que vous ne pouvez leur donner ou qui, selon vous, échappent à votre mandat	Y a-t-il des services particuliers demandés par les PME que vous ne pouvez pas leur fournir ou que, selon vous, ne cadre pas avec votre mandat ?
d. Until 2004, my parents met Nhan Thi Duong my ex-girlfriend and asked for my daughter Lan Thu Thi Le.	Ils n'ont rencontré Nhan Thi Duong, mon ex-petite amie, qu'en 2004, et lui ont demandé des nouvelles de ma fille, Lan Thu Thi Le.	Jusqu'en 2004, mes parents ont rencontré Nhan Thi Duong, mon ex-petite amie, et m'ont demandé de me donner ma fille Lan Thu Thi Le.
e. A cleaner, safer, more convenient road transportation system is possible – and closer to being realized than many believe.	Un système de transport routier plus propre, plus sûr et plus pratique est possible – et bien plus prêt de voir le jour qu'on ne le pense généralement. Il a seulement besoin de pouvoir faire ses preuves.	* Un système de transport routier plus propre, plus sûr et plus pratique est possible - et plus près d'être réalisé que beaucoup ne le croient.
f. A bigger bloodbath seems inescapable if he does not step down.	Il semble difficile d'échapper à un bain de sang plus important encore s'il n'accepte pas de démissionner.	* Un plus grand bain de sang semble inévitable s'il ne se retire pas.

Table 1: Examples of source sentences (SRC, first column), human (HUM, second column) and neural (NMT, third column) translations. Translations produced by DeepL are marked with a start, others are produced by in-house neural models. Errors are in bold and underlined, while bold highlights the concerned material

While it is difficult to appreciate why the classifiers we developed could accomplish such a distinction, we analyzed many NMT outputs and found among other things a strong tendency of systems to produce literal translations, mimicking the syntax of the source sentence. All examples in table 1 illustrate this tendency, as well as other typical cases we observed. First, systems have trouble with acronyms, as exemplified in example a) and b). Note that some acronyms are perfectly translated, as CHRC translated into CCDP in example b), but arguably, unless the system was exposed to such acronyms at training time (and assuming there is no ambiguity in their translation), there are no strong reasons to expect those units to be correctly (automatically) translated. We also occasionally found syntactical problems in automatic translations, such as example c) involving a failure in long-distance number agreement ("*cadre*" is wrongly generated in the singular form) as well as a wrong choice of pronoun ("*que*" instead of "*qui*"). Example d) showcases a situation where the trans-

lation went wrong in part because of the ambiguity of the phrase "ask for my daughter". Example f) shows a translation that our classifier could identify automatically, while we believe (with a non-expert eye), that the translation is correct. Replacing "*se retire*" by "*démissionne*" as a translation of "step down" reverses the classifier decision.

In the remainder, we measure the quality of various translation solutions with the so-called objective metric called BLEU (Papineni *et alii* 2002). We therefore hope that higher BLEU scores will be indicative of a better technology. We are aware that this score is nowhere near perfect, but it is nonetheless largely popular in the NLP community. Plus, the small-scale human evaluation we conducted turned out to be too complicated for our NLP team (which lacks a strong translation expertise) and far too time-consuming for this work. A least are we using BLEU consistently through the different experiments we report, therefore avoiding the problems discussed in Marie *et alii* (2021). Also, we have considered other objective metrics, with very similar outcomes overall.

3. Need for appropriate data

Even if technology has improved drastically, it still heavily relies on adequate parallel data to be trained on. By "adequate", we mean data representative of the targeted domain, as well as data with as few problems as possible, since the system might learn to reproduce them. We describe experiments we conducted with both issues.

3.1 Out-domain translation

In this section, we address the typical issue of adapting a system to a domain not well covered in the data used to train a neural engine with. This is a problem to anyone who plans to make good use of a generic engine to translate texts from a specific domain (e.g., medicine, finance). We conducted this work in collaboration with the *Autorité des marchés financiers du Québec* (AMF)[10], where our goal was to design an NMT engine for translating sentences of the financial domain. We first describe the main datasets available to train NMT engines on for the French-English language pair. Then, we report the main outcomes of our efforts in designing a system for the financial domain.

[10] https://lautorite.qc.ca/en/general-public (last consulted: June, 5th 2021).

The most known and well-studied English-French bilingual data is the training portion provided within the WMT'14 shared task (Bojar *et alii* 2014). It contains 40.8M sentence pairs extracted from five datasets that cover various domains: the Europarl V7 (Koehn 2005), the United Nations Corpus (Eisele, Chen 2010), the Common Crawl corpus[11], the News Commentary corpus, and the 10 French-English corpus[12]. Lison, Tiedemann (2016) present OpenSubtitles, a parallel corpus of 2.6 billion sentences across 60 languages originally mined from movies and television episode subtitles. The English-French portion of OPUS (Tiedemann 2012) contains a small collection (ECB) of parallel data in the financial domain, that is collected from documents published by the European Central Bank. Paracrawl[13] is an ongoing project aiming to collect parallel data from the web for all 24 official European Union languages. The dataset is extremely large and noisy, therefore there have been studies reported to filter in high quality pairs (Koehn *et alii* 2018). For example, even if the English-French subset contains over 4 billion sentence pairs, Ott *et alii* (2018) extracted 127M clean pairs after applying their filtering procedure.

Some statistics about those publicly available English-French corpora are given in table 2. Needless to say, for many language pairs, the parallel resources are much less abundant.

In Ghaddar, Langlais (2020), we describe efforts we made in order to acquire SEDAR, a large scale English-French parallel corpus for the financial domain. The first release of the corpus comprises 8.6 million high quality sentence pairs. The System for Electronic Document Analysis and Retrieval (SEDAR)[14] provides access to public security documents and information filed by Canadian issuers. The filings are

Dataset	Domain	# Sentence pairs (M)
Europarl	Politics	2.0
Common Crawl	Web	3.2
United Nations	Public	12.8
News Commentary	News	0.2
10^9 Word	General	22.6
OpenSubtitles	Movies	32.6
Opus-ECB	Finance	0.1
SEDAR (our work)	Finance	8.6

Table 2: Main characteristics of popular French-English publicly available parallel corpora, as well as SEDAR that we gathered (Figures are in millions)

[11] https://commoncrawl.org/the-data (last consulted: June, 5th 2021).
[12] https://www.statmt.org/wmt10/translation-task.html (last consulted: June, 5th 2021).
[13] https://paracrawl.eu (last consulted: June, 5th 2021).
[14] https://paracrawl.eu (last consulted: June, 5th 2021).

made available for personal and non-commercial use only, and it is strictly forbidden to extract them with an automatic process (e.g., a crawler)[15]. The data is the property of the Alberta Securities Commission on behalf of the Canadian Securities Administrators, the thirteen provincial and territorial Canadian securities regulatory authorities.

In collaboration with AMF, we created the SEDAR corpus, based on publicly-available documents and information filed in SEDAR between 1997 and 2018. Our point here is not to describe the many technological issues that we went through in preparing this dataset[16], but the importance of data. Our extensive MT experiments actually showed that SEDAR is essential to obtain good performance on the financial domain.

We split SEDAR in sedar-train (8.6M sentence pairs), sedar-dev (6k) and sedar-test (6k) datasets. The test set gathers data issued in 2018 (the last year of the data that was available at the time we mined the SEDAR website) ensuring that the sentences were not close to some training sentences (data up to 2017) or the validation material. In particular, we eliminate sentences that had more than 10% 4-grams in common with a sentence in the training/ development material.

We used the ConvS2S model as implemented in the fairseq toolkit with the same configuration that the authors used for the WMT'14 English-French experiments. We trained NMT models on subsets of size {2, 4, 6, 8} millions of randomly selected sentences pairs from sedar-train and WMT'14 (hereafter called WMT-RND). In addition, we experimented with SEDAR domain adapted subsets extracted from WMT'14 corpus, that we call WMT-SDA: we followed the approach of Alxelrod *et alii* (2011) and Moore, Lewis (2010) to select sentences from a large general domain parallel corpus (WMT'14) that are the most relevant to the target domain (finance). For this, we trained a 5-gram backoff language model (LM) using kenLM (Heafield 2011) on sedar-train in order to score WMT'14 sentences. We generated finance domain-adapted training data by selecting the highest scoring sentences (lowest perplexity). Finally, we also con-

[15] AMF will grant consulted to the SEDAR corpus without charge for academic research upon request˙ https://github.com/autorite/sedar-bitext (last consulted: September, 28th 2021).

[16] Converting pdf documents to text was to our surprise very difficult (financial texts contain many tables that confound/challenge? pdf converters). It took us roughly 2 months of a skilled computational engineer (first author) to address issues as various as: tokenization, sentence segmentation and alignment, and sentence deduplication. All those issues are detailed in Ghaddar, Langlais (2020).

sider systems trained on the sedar-train material directly. Results are reported in figure 1.

Expectedly, we observe that increasing the size of the training material impacts positively all the systems trained. It is obviously much more beneficial to train a system on the SEDAR material directly than using out-domain data. The best system we trained achieved a BLEU score of 38.3 on the test set, while the best system trained on the WMT material performed at best around 32. Also, we observe that selecting the out-domain data thanks to a language model (red over blue curve) is interesting, but that the positive effect eventually vanishes for larger training sets (8M). By inspecting LM scores, we noticed that at most the top 4 million pairs of WMT-SDA shares a high degree of similarity with SEDAR. The rest of sentences have similar low scores, which practically turns the selection to a random process.

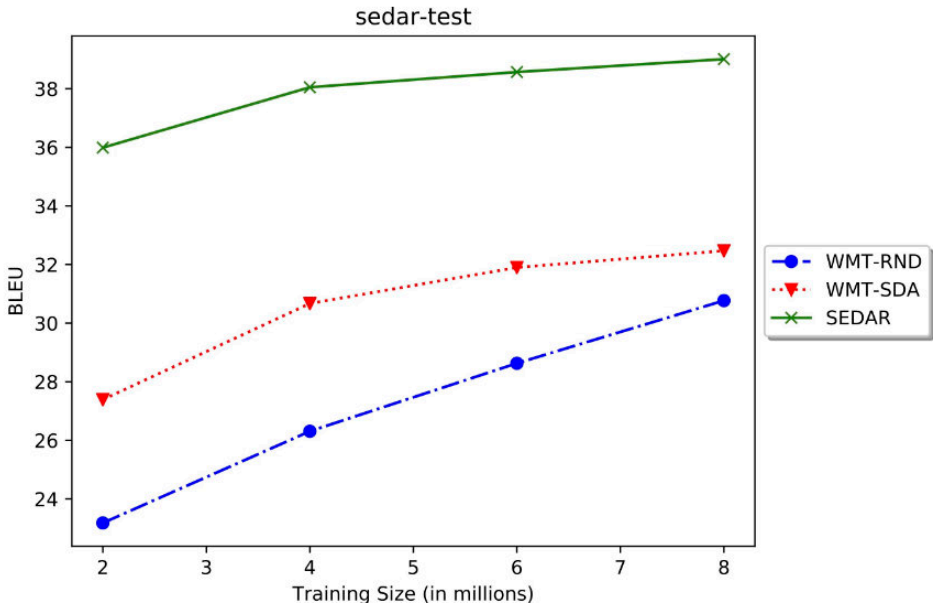


Figure 1: BLEU scores on sedar-test of models trained on increasing sized subsets of: sedar-train, WMT- RND, WMT-RND corpora. Read the text for more

We further experimented a setting where out-domain data is added to the training material of SEDAR, leading to a further BLEU gain of over 2 points. The large gain observed by using in-domain training data (SEDAR), possibly in conjunction with out-domain one (WMT) is not in itself surprising. Still, it allows us to pinpoint that gathering the appropriate data to train a system on is part of the MT ecosystem.

3.2 Multi-domain material

So far, we considered a situation typically studied in academic research: the one of adapting a translation engine to a specific domain. Here, we are considering a more practical setting where a system has to be used for translating data from many possibly overlapping domains. By "domain" here, we mean whatever metadata that accompanies subsets of a huge translation memory. We conducted these investigations in collaboration with the Translation Bureau of Canada (TBC) which granted us access to a huge translation memory (MC hereafter) gathering over 60 domains, that roughly correspond to specific areas of expertise such as Immigration (IMM), Geology (GEO), Mechanic (MEC), Fish Farming (AQU), Civil Engineering (CIV) or Administrative/ General Texts (TAG). Some domains are over-represented in the memory. For instance, one domain (TAG) comprises over 50% of the sentence pairs in the TM and concerns administrative and general texts, while 9 domains have less than 10k sentence pairs.

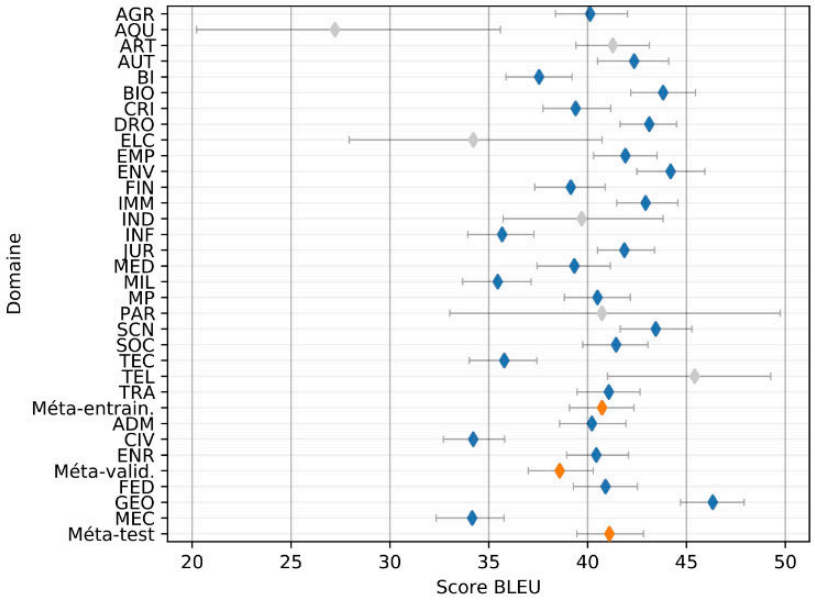


Figure 2: BLEU scores of the pre-trained model (WMT). Domains with less than 1600 sentence pairs are printed in grey. Domains are grouped into meta datasets: train (entrain), dev (valid) and test (test). Scores in orange are aggregated on each meta set. 95% confidence have been/was computed by bootstrap resampling

On test sets we gathered for each domain of interest, we measured the performance of the Scaling-NMT architecture, a transformer-based model that we found faster to train than the XLM one, while delivering slightly better BLEU scores. Scaling-NMT comes with a pretrained

model[17] available *via* the fairseq library. This model was trained on 35 million sentence pairs extracted from the WMT'14 dataset. Figure 2 shows the BLEU score of the system per domain.

We observe a difference of over 10 points between the domain with best BLEU scores (GEO) and the one with the lowest scores (MEC if we only consider test sets with at least 1 600 sentence pairs, AQU otherwise). This is a good illustration of what can be expected from a "generic" engine.

As a point of comparison, we also trained with fairseq a Scaling-NMT model on 19 million sentence pairs we gathered from the translation memory of the TBC, therefore an in-domain setting. The performances of this model are reported in figure 3. This system obtains much better BLEU scores than the pre-trained one, with the exception of the GEO domain, where the pre-trained system is at a slight (likely non-significant) advantage. This suggests that just collecting out-domain data is not enough to derive a system able to translate domain-specific texts. MEC and CIV remain the two domains for which both systems have the most difficulties. Therefore, it is not only a matter of throwing specific data of each domain of interest to a model for it to perform well on each domain it has seen.

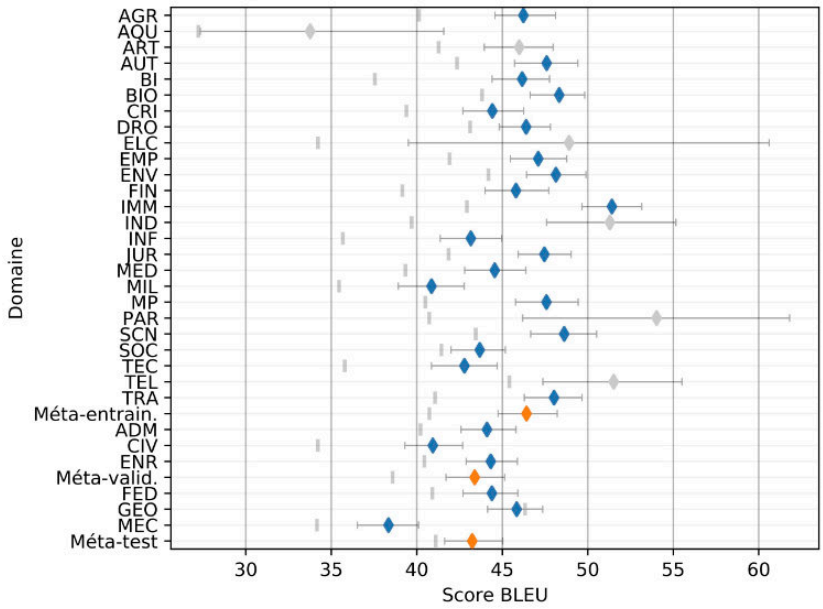


Figure 3: BLEU scores of the model trained on the 19M in-domain sentence pairs. Domains with less than 1 600 sentence pairs are printed in grey 95% confidence intervals have been computed by bootstrap resampling. Scores obtained by the pre-trained model (see Fig. 2) are indicated with a small vertical grey bar

[17] We used the transformer.wmt14.en-fr model.

So far, we have been using systems trained on a single training set, either the 35M sentence pairs of the WMT'14 dataset (out-domain, pre-trained), or the 19M sentence pairs we gathered from the TBC translation memory. Nowadays, fine-tuning a pre-trained system to a new task is an attractive solution (Devlin *et alii* 2019). In our case, fine-tuning a translation engine means continuing to update the parameters of a pre-trained model on a dataset (domain) of interest. There are several reasons for doing this. It might allow to adapt a model pre-trained on (possibly) out-domain data to a specific domain, by providing it data of this domain. Also, fine-tuning can be much less computation demanding than training a system from scratch. It may thus avoid making hard decisions about the training set to gather at an earlier stage of the development of a system.

We report on fine-tuning the two large models described previously with the goal of augmenting the performance on the MEC domain, which we purposely removed from the training material. According to pairwise domain similarity measurements, the closest domain to MEC in the translation memory is TEC, with 922 297 sentence pairs. Thus, we fine-tuned the two large systems we tested on TEC. It took 4.5 hours to fine-tune the WMT model, and slightly more for the TBC model, compared to approximatively 40 hours to train the system from scratch on the same GPU, demonstrating that fine-tuning is a reasonable option from a computational point of view.

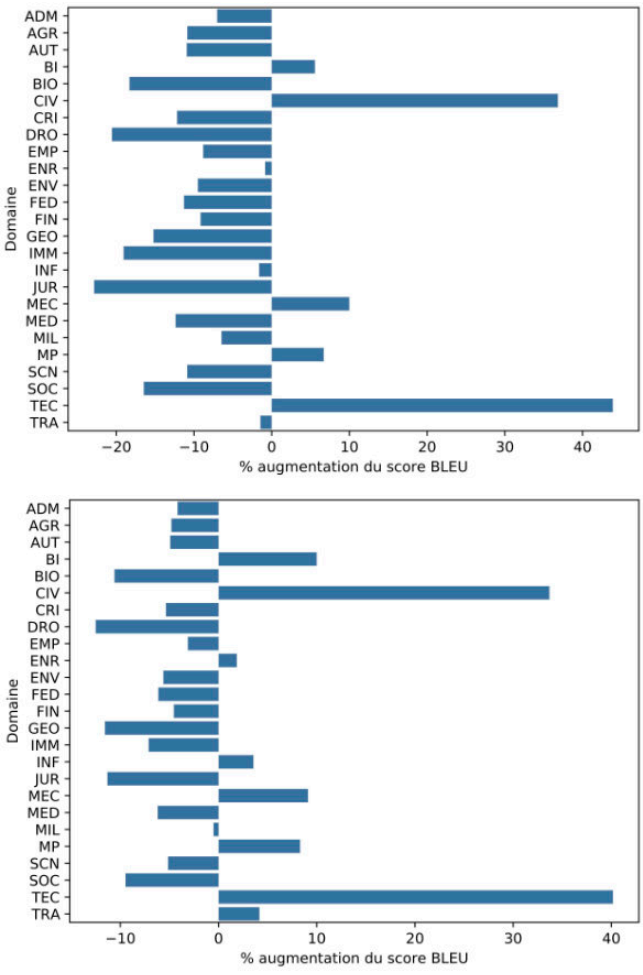


Figure 4: Gains of BLEU scores when fine-tuning the WMT model (up) and the in-domain model (down) on the TEC dataset

The relative gain of BLEU scores on the different domains are reported in figure 4. There are several things to be noted. First, the performance of both models on MEC increase by roughly 10%, which is satisfactory. Also, and without much surprise, the performance on the domain on which we fine-tune (TEC) received a huge boost of BLEU (40% or more for both systems). Recall that the TBC model has been also trained on TEC (among other domains), but that it still benefits from fine-tuning on TEC almost as much as the WMT model (which never saw TEC material). Last, we observe for both fine-tuned systems that the gains measured come with some degradation for a lot of other domains; the WMT model being the one which degrades the most. This suggests that in practice, one system should be fine-tuned for each domain of interest (or each group of domains), leading to as many systems as domains of interest.

We stress that many more experiments should be conducted for understanding the best practices for adapting a system to a specific domain, seen or not at (initial) training time. There are many parameters that might impact the performance of fine-tuning, including the size of the dataset used for fine-tuning. We might also leverage more than one close domain to improve the performance on a specific one. Also, many studies are currently being explored for adapting a system to a new domain, some of which we currently investigate.

We believe these observations contribute to show that even selecting material to train (from scratch or for finetuning) a translation system involves a great deal of choices that is definitely part of an NMT ecosystem, and which goes far above simply collecting the largest dataset possible.

3.3 Clean data

As we discussed, it is important to pay attention to the data we train a system on: gathering in-domain data is a key to develop a satisfactory translation engine. But even gathering appropriate data does not necessarily warrant the best system. On collaborating with the Translation Bureau of Canada, we studied the problem of cleaning their in-house translation memory. This organization is continuously feeding professional translations into a huge translation memory that is eventually used in their internal translation workbench. It turns out that even though their TM includes very high-quality translations, it does show some prob-

lems of various nature. For instance, file format conversion may corrupt sentences, which sometimes happens in a complex translation environment which accommodates many file formats (and translators). Sentence alignment errors might also occur when processing the texts to be added to the TM. We actually believe that most of the noise in this TM comes from the bad synergy between sentence segmentation and alignment. Some translations may also be of poor quality overall, although we did not notice such problems[18].

We studied the problem of cleaning a mostly clean professional TM and developed various technologies among which a set of various heuristics, feature based classifiers (SVM), deep learning ones (biLSTM), as well as the unsupervised multilingual-similarity search (MSS) algorithm from the LASER toolkit (Artetxe, Schwenk 2019) which uses language-agnostic pretrained sentence representations from 92 different languages (Schwenk *et alii* 2017). More details can be found in Bhardwaj *et alii* (2021b).

One challenge we found was to evaluate the cleaning technology we developed. As a matter of fact, in such a huge TM, we do not have the list of all the sentence pairs that might have a problem. We did inspect a sample of sentence pairs identified problematical by the technology we tested and realized how difficult it actually is to decide of the quality of a sentence pair. We can easily tell whenever an obvious error is present (such as a bad sentence alignment, or an encoding issue which compromises the source or the target material), but much less so for more subtle errors such as the omission (or insertion) of non-essential words. Instead, we resorted to using a neural translation engine as a proxy, with the idea that if a cleaning technology is efficient, then training a translation engine on the clean material should lead to a better BLEU score. And this, especially since we know NMT is affected by noise in the training material (Khayrallah, Koehn 2018).

Train	#SPs	XLM	ConvS2S
Baseline	14.5M	36.25	33.04
SVM	7.5M	36.53	33.91
Heuristics	8.2M	36.80	33.78
LASER	9.7M	37.23	33.58
biLSTM	6.1M	37.52	33.96
All	5.8M	37.57	33.93

Table 3: BLEU scores of two neural translation engines trained on different subsets of a bitext of 14.5M in-domain sentence pairs. Baseline corresponds to systems trained on the full bitext

[18] We did find a few sentence pairs with typos (source or target), as well as the presence of what we believe to be calques in certain translations.

For this, we gathered an in-domain bitext of 14.5M sentence pairs that we used for training a baseline system. We then cleaned out this material by applying our cleaning technology, and trained systems on the resulting bitexts. All the systems were then used to translate a representative test set and the translations evaluated with BLEU. The main outcome of this is reported in Table 3.

We used two types of models to ensure that our observations were not specific to a given system: the transformer-based system XLM and the recurrent network ConvS2S, the former outperforming the latter in our experiments. What is more interesting is what happens when deploying the cleaners we developed. For instance, using our set of heuristics, we removed almost half of the sentence pairs of our bitext without any negative noticeable impact on BLEU. The most effective cleaner is actually the deep learning classifier we trained (biLSTM) that removed more than half of the material with noticeable improvements in terms of BLEU. Combining all the cleaners (last line) leads to an even better setting: less training material, and better BLEU scores.

Arguably, using machine translation is not entirely satisfying here, and there are many reasons that could explain these results. First, the test material might be specific, and the cleaners could simply operate data selection. Also, as we discussed already, BLEU might have some undesired biases. Last, by inspection, we found many sentence pairs filtered that did seem of good quality to us. All these threats to validity are further analyzed in Bhardwaj *et alii* (2021b). Still, we believe those results suggest that cleaning data is an important issue which is not well studied, especially when the data is expected to be of good quality, as was our case here.

4. Discussion

As we have been discussing through this article, NMT technology has limitations, and a great deal of experiments are required in order to find the appropriate architecture to use, to prepare the right data for a given task, and to devise good strategies for adapting an engine to a specific use case. In our laboratory (RALI), we conducted all the experiments we reported in roughly a 2-year period, involving skilled NLP students, well trained to deep learning platforms. Of course, if we had to do it again, we

would definitely do it faster. Still, we believe that unless an organization delegates all this expertise to a trustworthy company specialized in adapting MT technology, there is no choice than to employ NLP experts to maintain a translation solution of quality.

Technological expertise is definitely one asset, but there are many other things in an MT ecosystem beyond this. We are strongly convinced that a fruitful use of neural MT necessitates the maturation of appropriate user interfaces. Post-editing the output of machine translation is certainly a solution, but arguably, finding errors in current MT output might be more difficult since translations are fluent. We already discussed works that suggest that NMT leads — once post-edited — to better translation quality, but more such studies should be conducted. Meanwhile, technology for spotting potential problems in the output of NMT might prove useful in practice.

Also, one should think of efficient ways of integrating MT technology into the translation environment the professional translator has access to. Interactive machine translation (IMT), where a system interacts with a professional translator to produce a translation, might be a more useful way to rapidly deliver high quality translations. IMT (Foster *et alii* 2002; Casacuberta *et alii* 2009) were proposed when so-called statistical translation came on the scene, and well-designed interfaces were proposed, see for instance the Casmacat project (Alabau *et alii* 2014a) in which e-pen interaction, as well as hand-written recognition (Alabau *et alii* 2014b) have been integrated. IMT built on top of neural translation is nowadays blooming (Grangier, Auli 2018; Sebastin *et alii* 2019; Wang *et alii* 2020), and Knowles *et alii* (2019) reports that half the translators that tested their prototype were translating faster than post-editing.

Last, in all the experiments reported, we have been using so-called objective metrics to measure translation quality. While they are good metrics to gear specific technological developments, we had many occasions to observe that they are only crude ways of measuring quality. We are convinced that a good MT ecosystem must involve repeated human expert evaluations, monitoring specific activities conducted by professional translators in their typical production environment. Only then will we find the proper ways of using MT efficiently.

In doing so, we might learn that NMT is one asset, but among many others. A professional translator is accomplishing many activities, among

which consulting dedicated applications over the Web, such as bilingual concordancers (e.g., Linguee[19] or TransSearch (Bourdaillet *et alii* 2010)), terminological databases (e.g., Termium[20]) or in-house translation memories. We believe that a harmonized way of consulting all these useful sources of information would, in conjunction with MT post-edition or interactive machine translation, lead to a fruitful synergy.

Conclusion

In this paper, we have been discussing issues involved in developing neural machine translation, using three popular and efficient libraries. We made clear that such an endeavour requires a mix of expertise, including computer science, a familiarity with deep learning architectures, and a real care for data. This suggests that NMT is for many organizations a service that must be acquired from a third-party company. But one must still realize that this company should provide adequate facilities for training/adapting a system to a specific need or data: generic (N)MT is not enough in practice (at least this is what we observed). We also warn that despite impressive results, neural MT is typically not delivering human-like quality. Translations are often literal, and may contain problems that are perhaps less easy to spot due to the overall fluency of the output produced. While there are situations where this might be good enough, quality translation is far not a solved problem, and it is — we believe — dangerous to think so.

[19] https://www.linguee.com

[20] https://www.btb.termiumplus.gc.ca/

References

Alabau Vicent, Buck Christian, Carl Michael, Casacuberta Francisco, García-Martínez Mercedes, Germann Ulrich, González-Rubio Jesús, Hill Robin, Koehn Philipp, Leiva Luis, Mesa-Lao Bartolomé, Ortiz-Martínez Daniel, Saint-Amand Hervé, Sanchis Trilles Germán, Tsoukala Chara (2014a). "CASMACAT: A Computer-assisted Translation Workbench". In: Shuly Wintner, Marko Tadić, Bogdan Babych (eds). *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 25-28.

Alabau Vicent, Sanchis Alberto, Casacuberta Francisco (2014b). "Improving on-line handwritten recognition in interactive machine translation". *Pattern Recognition*, 47/3, 1217-1228.

Artetxe Mikel, Schwenk Holger (2019). "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". *Transactions of the Association for Computational Linguistics,* 7, 597-610.

Axelrod Amittai, He Xiaodong, Gao Jianfeng (2011). "Domain Adaptation via Pseudo In-Domain Data Selection". In: Regina Barzilay, Mark Johnson (eds). *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK., 355-362.

Bahdanau Dzmitry Cho Kyung Hyun, Bengio Yoshua (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". *ArXiv*, abs/1409.0473. https://arxiv.org/abs/1409.0473 (Last consulted: June 10th 2021).

Bhardwaj Shivendra, Alfonso Hermelo David, Langlais Philippe, Bernier-Colborne Gabriel, Goutte Cyril, Simard Michel (2021a). "Human or Neural Translation?". In: Donia Scott, Nuria Bel, Chengqing Zong (eds). *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain, 6553-6564.

Bhardwaj Shivendra, Alfonso Hermelo David, Langlais Philippe, Bernier-Colborne Gabriel, Goutte Cyril, Simard Michel (2021b). "Cleaning a(n almost) Clean Institutional Translation Memory". In: *Technical Report*, University of Montreal.

Bojar Ondřej, Buck Christian, Federmann Christian, Haddow Barry, Koehn Philipp, Leveling Johannes, Monz Christof, Pecina Pavel, Post Matt, Saint-Amand Hervé, Soricut Radu, Specia Lucia, Tamchyna Aleš (2014). "Findings of the 2014 Workshop on Statistical Machine Translation". In: Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Lucia Specia (eds). *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA, 12-58.

Bourdaillet Julien, Huet Stéphane, Langlais Philippe, Lapalme Guy (2010). "TransSearch: from a bilingual concordancer to a translation finder". *Machine Translation*, 24/3, 241-271.

Brown Peter, Della Pietra Stephen, Pietra, Vincent, Mercer Robert (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation". *Computational Linguistics*, 19/2, 263-311.

Casacuberta Francisco, Civera Jorge, Cubel Elsa, Lagarda Antonio L., Lapalme Guy, Macklovitch Elliott, Vidal Enrique (2009). "Human interaction for high-quality machine translation". *Communications of the ACM*, 52/10, 135-138.

Castilho Sheila, Moorkens Joss, Gaspari Federico, Calixto Iacer, Tinsley John, Way Andy (2017). "Is Neural Machine Translation the New State of the Art?". *The Prague Bulletin of Mathematical Linguistics*, 108, 109-120.

Conneau Alexis, Lample Guillaume (2019). "Cross-lingual Language Model Pretraining". In: Hanna Wallach and Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, Roman Garnett (eds). *Proceedings of the 33rd Conference on Neural Information Processing Systems*, Vancouver, Canada, 7059-7069.

Dauphin Yann N., Fan Angela, Auli Michael, Grangier David (2017). "Language Modeling with Gated Convolutional Networks". In: Doina Precup, Yee Whye Teh (eds). *Proceedings of the 34th International Conference on Machine Learning,* Sydney, Australia, 933-941.

Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: Jill Burstein, Christy Doran, Thamar Solorio (eds). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language,* Minneapolis, Minnesota, USA, 4171-4186.

Escribe Marie (2019). "Human Evaluation of Neural Machine Translation: The Case of Deep Learning". In: Irina Temnikova, Constantin Orasan, Gloria Corpas Pastor, Stephan Vogel (eds). *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology*, Varna, Bulgaria, 36-46.

Eisele Andreas, Chen Yu (2010). "MultiUN: A Multilingual Corpus from United Nation Documents". In: Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, Daniel Tapias (eds). *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta, 2868-2872

Foster George, Langlais Philippe, Lapalme Guy (2002). "TransType: Text Prediction for Translators". In: Mitchell Marcus (ed). *Proceedings of the second international conference on Human Language Technology Research (Demonstrations),* San Diego California, USA, 372-374.

Gehring Jonas, Auli Michael, Grangier David, Yarats Denis, Dauphin Yann N. (2017). "Convolutional sequence to sequence learning". In: Doina Precup, Yee Whye Teh (eds). *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 1243-1252.

Ghaddar Abbas, Langlais Philippe (2020). "SEDAR: a Large-Scale French-English Financial Domain Parallel Corpus". In: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds). *Proceedings of the 12th International Conference on Language Resources and Evaluation*, Marseille, France, 3595-3602.

Goodfellow Ian, Bengio Yoshua, Courville Aaron (2016). *Deep Learning*, Cambridge, MIT Press.

Grangier David, Auli Michael (2018). "QuickEdit: Editing Text & Translations by Crossing Words Out". In: Marilyn Walker, Heng Ji, Amanda Stent (eds). *Proceedings of 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, New Orleans, Louisiana, USA, 272-282.

Hassan Hany, Aue Anthony, Chen Chang, Chowdhary Vishal, Clark Jonathan, Federmann Christian, Huang Xuedong, Junczys-Dowmunt Marcin, Lewis William, Li Mu, Liu Shujie, Liu, Tie-Yan, Luo Renqian, Menezes Arul, Qin Tao, Seide Frank, Tan Xu, Tian Fei, Wu Lijun, Zhou Ming (2018). "Achieving Human Parity on Automatic Chinese to English News Translation". *ArXiv*, abs/1803.05567, https://arxiv.org/abs/1803.05567 (Last consulted: June 5th, 2021).

Heafield Kenneth (2011). "KenLM: Faster and Smaller Language Model Queries". In: Chris Callison-Burch, Philipp Koehn, Christof Monz, Omar F. Zaidan (eds). *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, 187-197.

Isabelle Pierre, Cherry Colin, Foster George (2017). "A Challenge Set Approach to Evaluating Machine Translation". In: Martha Palmer, Rebecca Hwa, Sebastian Riedel (eds). *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2486-2496.

Khayrallah, Huda, Philipp Koehn (2018). "On the impact of various types of noise on neural machine translation". In: Alexandra Birch, Andrew Finch, Thang Luong, Graham Neubig, Yusuke Oda (eds). *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia, 74-83.

Knowles Rebecca, Sanchez-Torron Marina, Koehn Philipp (2019). "A user study of neural interactive translation prediction". *Machine Translation,* 33, 135-154.

Koehn Philipp (2005). "Europarl: A parallel corpus for statistical machine translation". In: Yaser Al-Onaizan, Will Lewis (eds). *Proceedings of the Tenth Machine Translation Summit*, Miami, Florida, USA, 79-86.

Koehn Philipp, Khayrallah Huda, Heafield Kenneth, Forcada Mikel L. (2018). "Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering". In: Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Gra-

ham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, Karin Verspoor (eds). *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Brussels, Belgium, 726-739.

Lison Pierre, Tiedemann Jörg (2016). "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles". In: Portorož, Slovenia, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds). *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 923-929.

Lopes António, Farajian M. Amin, Bawden Rachel, Zhang Michael, Martins André F. T. (2020). "Document-level Neural MT: A Systematic Comparison". In: André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, Mikel L. Forcada (eds). *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal, 225-234.

Marie Benjamin, Fujita Atsushi, Rubino Raphael (2021). "Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers". In: Chengqing Zong, Fei Xia, Wenjie Li, Roberto Navigli (eds). *Proceedings of the joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, 7297-7306.

Moore Robert C., Lewis William (2010). "Intelligent Selection of Language Model Training Data". In: Jan Hajič, Sandra Carberry, Stephen Clark, Joakim Nivre (eds). *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics Conference*, Uppsala, Sweden, 220-224.

Ott Myle, Edunov Sergey, Grangier David, Auli Michael (2018). "Scaling Neural Machine Translation". In: Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, Karin Verspoor (eds). *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium, 1-9.

Ott Myle, Edunov Sergey, Baevski Alexei, Fan Angela, Gross Sam, Ng Nathan, Grangier David, Michael Auli (2019). "Fairseq: A fast, extensible toolkit for sequence modeling". In: Waleed Ammar, Annie Louis, Nasrin Mostafazadeh (eds). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, USA, 48-53.

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

Samuel Läubli, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, Martin Volk (2019). "Post-editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain". In: Mikel Forcada, Andy Way, Barry Haddow, Rico Sennrich (eds). *Proceedings of Machine Translation Summit XVII: Research Track*, Dublin, Ireland, 267-272.

Sánchez-Gijón Pilar, Moorkens Joos, Way Andy (2019). "Post-editing neural machine translation versus translation memory segments". *Machine Translation,* 33/1-2, 1-29.

Santy Sebastin, Dandapat Sandipan, Choudhury Monojit, Bali Kalika (2019). "IN-MT: Interactive Neural Machine Translation Prediction". In: Sebastian Padó, Ruihong Huang (eds). *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations*, Hong Kong, China, 103-108.

Schwenk Holger, Tran Ke, Firat Orhan, Douze Matthijs (2017). "Learning Joint Multilingual Sentence Representations with Neural Machine Translation". In: Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, Scott Yih (eds). *Proceedings of the 2nd Workshop on Representation Learning for NLP*, Vancouver, Canada, 157-167.

Sennrich Rico, Haddow Barry, Birch Alexandra (2016). "Neural Machine Translation of Rare Words with Subword Units". In: Katrin Erk, Noah A. Smith (eds). *Proceedings of the 54th Annual Meeting for the Association of Computational Linguistics*, Berlin, Germany, 1715-1725.

Statt Nick (2016). "Google's AI translation system is approaching human-level accuracy". Blog. https://www.theverge.com/2016/9/27/13078138/google-translate-ai-machine-learning-gnmt (Last consulted: July 14th, 2021).

Sutskever Ilya, Vinyals Oriol, Le Quoc V. (2014). "Sequence to sequence learning with neural networks". In: Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, Kilian Q. Weinberger (eds). *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal Canada, 3104-3112.

Tiedemann Jörg (2012). "Parallel Data, Tools and Interfaces in OPUS". In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds). *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2214-2218.

Toral Antonio, Castilho Sheila, Hu Ke, Way Andy (2018a). "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation". In: Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, Karin Verspoor (eds). *Proceedings of the Third Conference on Machine Translation: Research Papers,* Brussels, Belgium, 113-123.

Toral Antonio, Wieling Martijn, Way Andy (2018b). "Post-editing Effort of a Novel with Statistical and Neural Machine Translation". *Frontiers in Digital Humanities,* 5, 2297-2668.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Łukasz, Polosukhin Illia (2017). "Attention is All you Need". In: Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, Rob Fergus (eds). *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, California, USA, 5998-6008.

Wang Qian, Zhang Jiajun, Liu Lemao, Huang Guoping, Zong Chengqing (2020). "Touch Editing: A Flexible One-Time Interaction Approach for Translation". In: Suzhou, China, Kam-Fai Wong, Kevin Knight, Hua Wu (eds). *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 1-11.

Xu Jitao, Josep Crego, Jean Senellart (2020). "Boosting Neural Machine Translation with Similar Translations". In: Dan Jurafsky, Joyce Chai, Natalie Schluter, Joel Tetreault (eds). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 1580-1590.

Zhou Sharon (2018). "Has AI surpassed humans at translation? Not even close!". Blog. http://www.skynettoday.com/editorials/state_of_nmt (Last consulted: July 14th, 2021).

Deuxième partie : expérimentations pédagogiques

De Europa Deuxième partie - Special Issue 2022

Introduction
Expérimentations pédagogiques : perception et utilisation de l'intelligence artificielle dans la formation universitaire*

Alida Maria Silletti, Rachele Raus

Dans cette partie de l'ouvrage, il s'agit de présenter l'enquête menée en salle de classe universitaire en France et en Italie par des questionnaires que plusieurs personnes des groupes de travail italophone et francophone ont voulu soumettre à leurs étudiantes et à leurs étudiants au cours de l'année universitaire 2020-2021 (premier et second semestre) au sujet de l'intelligence artificielle.

Dans le tableau 1, il est possible de voir plusieurs données intéressantes par rapport aux deux questionnaires de début (Q1) et de fin de cours (Q2) qui ont été soumis, notamment :

1) le nombre des personnes (licence, M1 ou M2) auxquelles les deux questionnaires, ou, dans quelques cas, seulement le premier, ont été soumis. Au total, le premier questionnaire a été rempli par 555 personnes et le second par 299. Dardo de Vecchi a soumis un seul questionnaire (tous les niveaux et programmes confondus) qui a été élaboré à partir du questionnaire-type soumis plus généralement par les autres professeurs et professeures qui ont participé à l'initiative. Ce questionnaire a été rempli par 609 personnes ;
2) la période de soumission, qui a tenu compte des calendriers académiques des institutions concernées en France et en Italie. La durée de soumission a été d'environ 3-5 jours, sauf pour le questionnaire de Dardo de Vecchi qui a été soumis entre le 5 mars et le 21 avril 2021. On a envoyé un lien utile aux participantes et aux participants, qui ont pu remplir leur questionnaire sous forme anonyme par le modèle *Google form* une seule fois ;

Alida Maria Silletti, Università di Bari, alida.silletti@uniba.it
Rachele Raus, Università di Bologna, rachele.raus@unibo.it
* Les paragraphes 1.1. et 1.2 ont été rédigés par Alida Maria Silletti. Les conclusions ont été écrites ensemble par les deux auteures.

3) le type de questionnaire soumis, en sachant que ceux qui ont été utilisés en Italie (IT) et en France (FR) ont été de deux types : questionnaire de base ou « type » et questionnaire avec des questions ajoutées (AD). Par rapport à ces derniers, les questionnaires italiens ont prévu l'ajout au questionnaire-type de questions supplémentaires concernant les genres textuels (voir Cennamo, Mattioda ; Cinato ; Molino). Par contre, le seul questionnaire français qui a prévu l'ajout de questions spécifiques a été celui qui a été soumis par Dardo de Vecchi, qui paraît au fond de son article. Les autres questionnaires ont été insérés comme annexes au fond de cet ouvrage.

Questionnaires soumis par	Q1	Nombre de personnes participantes	Q2	Nombre de personnes participantes	Type de questionnaire
Altmanova Jana Université L'Orientale de Naples	01/03/21	36 Master 2	∅	∅	Q1_IT Q2_IT
Altmanova Jana Université L'Orientale de Naples	02/03/21	44 Master 2	∅	∅	Q1_IT Q2_IT
Bisiani Francesca Université Catholique de Lille	07/01/21	18 Master 1	01/04/21	18	Q1_FR Q2_FR
Bisiani Francesca Université Catholique de Lille	07/01/21	20 Master 2	01/04/21	18	Q1_FR Q2_FR
Cinato Lucia Université de Turin	21/02/21	68 Licence	21/04/21	30	Q1_IT Q2_IT_AD
Dankova Klara Université Catholique de Milan	01/03/21	4 Master 1	02/05/21	2	Q1_IT Q2_IT
Dankova Klara Université Catholique de Milan	01/03/21	7 Licence	02/05/21	4	Q1_IT Q2_IT
de Vecchi Dardo Kedge Business School et alii	05/03/21	609 Tous niveaux confondus	∅	∅	Q1_FR_AD
Gournay Lucie Université de Paris 12	01/02/21	30 Licence	∅	∅	Q1_FR Q2_FR
Gournay Lucie Université de Paris 12	01/02/21	3 Master	∅	∅	Q1_FR Q2_FR
Mattioda Maria Margherita, Cennamo Ilaria Université de Turin	21/02/21	43 Licence	21/04/21	31	Q1_IT Q2_IT_AD
Mattioda Maria Margherita, Cennamo Ilaria Université de Turin	21/03/21	19 Licence	∅	∅	Q1_IT Q2_IT_AD
Molino Alessandra Université de Turin	01/03/21	125 Master 2	01/03/21	128	Q1_IT Q2_IT_AD
Raus Rachele Université de Turin	01/03/21	78 Master 1	01/04/21	68	Q1_IT Q2_IT
Zanola Maria Teresa Université Catholique de Milan	01/03/21	11 Master	∅	∅	Q1_IT Q2_IT
Zanola Maria Teresa Université Catholique de Milan	02/03/21	12 Master	∅	∅	Q1_IT Q2_IT
Zollo Silvia Domenica, Calvi Silvia Université de Vérone	21/02/21	26 Licence	∅	∅	Q1_IT Q2_IT
Zollo Silvia Domenica, Calvi Silvia Université de Vérone	21/02/21	11 Licence	∅	∅	Q1_IT Q2_IT

Tableau 1 : Questionnaires soumis pendant l'année universitaire 2020-2021

Nous faisons remarquer que les personnes qui ont soumis les questionnaires n'ont pas forcément participé à cette publication. Par contre, comme le projet a prévu l'adoption de la méthode de recherche-action, les participantes et les participants ont pu jouer à leur tour un rôle actif, en proposant des travaux et même des contributions qui ont pu être publiés, ou en tous les cas, intégrés dans cet ouvrage (voir l'article d'Abbadessa, Albini, De Paoli, Del Nobile, ainsi que la participation de Bottiglieri à l'article d'Altmanova).

1. L'élaboration et les résultats des questionnaires

Les questionnaires ont été élaborés à partir du modèle proposé par Zoltán Dörnyei (2007 : 102) dans les champs de la psycholinguistique et de la didactique. D'après ce modèle, les questions ont été subdivisées dans les trois groupes suivants : « *factual questions* », « *behavioural questions* », « *attitudinal questions* ». Le premier bloc vise à déterminer le profil de la personne à laquelle on soumet le questionnaire, le deuxième permet de s'informer sur la connaissance générale relative à l'intelligence artificielle (IA) et sur l'usage des dispositifs qui s'appuient sur des algorithmes d'apprentissage profond et le troisième est dédié aux opinions sur l'usage des outils terminologiques et de traduction automatique (TA).

Lorsqu'on a pu soumettre les deux questionnaires, la comparaison des données recueillies a permis d'évaluer les compétences acquises par les personnes qui ont participé à l'initiative.

Le questionnaire proposé par Dardo de Vecchi, qui ciblait un public un peu différent par rapport aux autres, a été élaboré à partir des questionnaires du projet mais en prenant en compte aussi les spécificités du public concerné (voir dans son article).

Les langues d'administration des questionnaires ont été l'italien (Q_IT) pour le public italien et le français (Q_FR) pour le public en France.

1.1 Résultats des premiers questionnaires

La population à laquelle ont été soumis les questionnaires se compose pour la plupart d'étudiantes[1], comme il était prévisible en raison du type d'études effectuées, qui sont en langues étrangères appliquées et en

[1] Elles sont majoritaires tant en licence qu'en master, sauf pour de Vecchi.

langues et littératures étrangères, à deux exceptions près (voir *infra*). Cette population est âgée de 19 à 38 ans — la plupart appartient à la tranche d'âge 19-22 ans en licence et 22-25 ans en master — et possède, dans la très grande majorité des cas, une maîtrise de l'italien (participantes et participants étudiant en Italie) et du français (participantes et participants étudiant en France) au niveau de langue maternelle. Les autres langues maternelles différentes de l'italien (en Italie) et du français (en France) relèvent de langues de l'immigration et se répartissent de manière non homogène dans les établissements universitaires des deux pays.

En Italie, parmi les langues maternelles différentes de la langue nationale, signalons le roumain, l'arabe (décliné sous ses différentes variétés), l'albanais, le polonais, le bosniaque et l'allemand. On peut également relever la présence de langues maternelles liées à la mobilité internationale : tel est le cas des étudiantes qui ont le français comme langue maternelle et qui ont répondu au questionnaire pendant leur mobilité Erasmus+ en Italie (voir Altmanova, Bottiglieri). Concernant les langues étrangères connues par la population installée en Italie, au-delà de l'anglais, qui semble être maîtrisé par tout le monde, quoique avec des niveaux différenciés allant du B1 au C1-C2 du CECR, la plupart des participantes et des participants déclarent connaître le français, l'allemand[2], l'espagnol et, par ordre décroissant, d'autres langues européennes et extra-européennes[3].

En France, l'administration du questionnaire a concerné deux cas non homogènes. D'une part, il s'agit d'étudiantes et étudiants en droit (voir Bisiani) de Master 1 et 2, qui sont surtout de langue maternelle française. Au sein de ce groupe, des personnes déclarent être de langue maternelle anglaise, espagnole, italienne, ghomala/ewondo ou thaïlandaise. Les étudiantes et les étudiants possèdent normalement une maîtrise d'au moins deux langues distinctes de leur langue maternelle. D'autre part, le questionnaire administré en France par de Vecchi concerne un cas *sui generis* : en effet, l'auteur a soumis un questionnaire partiellement différent du ques-

[2] Nous tenons à préciser que ces résultats sont en partie liés aux cours de langue étrangère (de français, d'anglais et d'allemand) au sein desquels les questionnaires ont été soumis.

[3] La connaissance de langues non européennes, notamment liées à l'arabe ou bien aux langues asiatiques, s'explique par deux raisons principales : d'un côté, certains cours de langue s'inscrivent dans des licences ou masters qui prévoient également l'enseignement de langues non romanes et non germaniques. De l'autre, c'est toujours le contexte migratoire d'origine des étudiantes et des étudiants (ou de leurs familles) qui explique leur maîtrise de langues non apprises dans le contexte scolaire ou universitaire, tous groupes confondus.

tionnaire-type de départ à 609 personnes inscrites à cinq écoles d'économie réparties sur le territoire français métropolitain. La population répondant à ce questionnaire est pour la plupart masculine. Ce dernier questionnaire est donc différent des autres aussi par rapport au public ciblé.

En effet, les autres questionnaires qui ont été administrés en France comme en Italie l'ont été dans le cadre des cours de langue étrangère (de français, d'anglais, d'allemand) sur objectifs spécifiques ciblant des compétences ainsi que des sujets différents. Dans la plupart des cas, ces cours ont visé le travail de traduction spécialisée et/ou la terminologie, appliquées à des domaines différents : du texte touristique au texte promotionnel en passant par le texte juridique et informatif.

La première section des questionnaires a permis de vérifier l'intérêt et les connaissances préliminaires à l'égard de l'IA à des fins didactiques. Les participantes et les participants étant des personnes qui étudient les langues étrangères, ou qui se servent d'une ou plusieurs langues étrangères à des fins professionnelles comme dans le cas des questionnaires soumis par de Vecchi, elles et ils emploient normalement des outils issus de l'IA, mais de manière peu consciente. C'est pourquoi l'un des objectifs du questionnaire était de sensibiliser cette population à l'emploi méthodique de ces outils.

Ainsi les premières questions des questionnaires ont-elles permis de vérifier l'intérêt des participants et des participantes à l'égard de l'IA au début du cours de langue pour ensuite vérifier, pour celles et ceux qui ont soumis le second questionnaire, si cet intérêt avait changé pendant et en fin de cours, et si oui, de quelle manière[4]. La plupart des participantes et des participants aux enquêtes ont déclaré être intéressées et intéressés aux outils de l'IA, bien qu'avec des pourcentages différenciés. En tous les cas, les réponses, tous profils confondus, montrent un intérêt « très » ou « assez » élevé pour ces outils.

Afin de vérifier les connaissances préliminaires des étudiantes et des étudiants, une question dirigée, mais prévoyant également des réponses ouvertes sur les plateformes technologiques déjà connues et/ou utilisées à des fins universitaires, a mis en évidence que cette population connaît bien les réseaux sociaux, notamment Facebook, Instagram et

[4] Les participantes et les participants ont eu à disposition 3-5 jours pour remplir et envoyer leurs questionnaires. Seulement de Vecchi a laissé ces derniers ouverts pendant toute la durée du cours.

WhatsApp, auxquels s'ajoutent les plateformes didactiques liées au contexte dans lequel les questionnaires ont été soumis. Il n'est pas étonnant de trouver Microsoft Teams, Moodle, Zoom et Webex parmi les outils technologiques connus par cette population. La liste précédente est complétée par des dictionnaires et des portails en ligne, parmi lesquels une distinction préliminaire n'est pas toujours faite entre des dictionnaires traditionnels, disponibles aussi en version numérique, voire « numériquée » (Paveau 2017), dans le réseau, comme le dictionnaire de la langue italienne Treccani, et des plateformes de traduction en ligne permettant d'accéder également à un dictionnaire, comme ReversoContext.

D'autres données qui émergent des premiers questionnaires portent sur le rôle de l'IA dans la poursuite des études et dans l'avenir professionnel des étudiantes et des étudiants. Les réponses à ces deux questions sont similaires : dans un avenir proche, à savoir après la fin des études, l'impact de l'IA est considéré comme élevé, voire maximal, au sujet de son utilité didactique tandis que son utilisation est mise en question par rapport à l'avenir professionnel. Surtout si elle est inscrite en licence, cette population n'est pas encore certaine des choix professionnels futurs, d'autant plus que le contexte pandémique et post-pandémique ne permet pas de regarder de manière « assurée » vers l'avenir. Cela dit, il faut également préciser que cette population ne connaît pas encore le potentiel de l'IA dans tous les domaines professionnels, d'où ses doutes.

Ces questions préliminaires sont suivies d'une section concernant la TA et ses outils. Une première remarque, qui peut sans doute conforter le travail des enseignantes et des enseignants de langues étrangères et de traduction, concerne la fiabilité qui est accordée aux outils de l'IA pour la TA montrée par les résultats du premier questionnaire. Très peu d'étudiantes et étudiants (tous parcours confondus) considèrent ces outils comme très fiables ou, au contraire, comme nullement fiables, la plupart (75 % des cas en moyenne) leur attribuant une fiabilité assez élevée. Cette population déclare effectuer des recherches supplémentaires pour vérifier les solutions proposées par les traducteurs automatiques : elle est donc amenée à analyser de manière critique les résultats de ces derniers, comprenant que seule l'intervention humaine permet de restituer une solution de traduction fiable. Parmi les outils connus par cette population, les plus utilisés sont Google Traducteur, ReversoContext,

DeepL et WordReference, dont elle tend à se servir surtout pour des traductions de et vers l'anglais, ensuite vers le français ou l'allemand[5].

Quant à la manière dont cette population perçoit de prime abord les résultats issus des outils de TA, la plupart des personnes considèrent que l'inadéquation de la solution proposée est due à l'idiomaticité ou à la complexité de l'expression à traduire, qui poseraient des difficultés à l'outil de TA. Selon les participantes et les participants, il est aussi possible que l'outil de TA soit inadéquat car son entraînement ne porterait pas sur le type de corpus ou de texte auquel appartient le mot ou l'expression à traduire. La population ciblée est donc consciente des limites des outils de TA, indépendamment des études qu'elle est en train d'effectuer en langues ou en droit et du pays concerné. Une question qui a permis de mieux comprendre les mécanismes qui font que les étudiantes et les étudiants privilégient un mot ou une expression proposés par un outil de TA concernait la vérification des sources de la TA, ce qui a montré rarement la capacité de s'interroger sur ces dernières. Les résultats laissent apercevoir, en gros, un degré de confiance non négligeable vis-à-vis des outils bilingues et multilingues de TA.

Comme l'enquête a porté sur le multilinguisme et que les étudiantes et les étudiants connaissent et/ou sont en train d'apprendre plus d'une langue étrangère, il était essentiel de comprendre si le degré de fiabilité des résultats de la TA était lié à la langue ciblée. La langue anglaise et plusieurs langues romanes apparaissent comme celles sur lesquelles les outils de TA travaillent le mieux, au contraire de l'allemand ou de l'espagnol, dont la traduction automatique est perçue comme plus littérale. Les participantes et les participants ont normalement donné des explications « techniques » des résultats de performance de traduction automatique obtenus pour les différentes langues concernées. Tel est le cas, par exemple, des langues sémitiques, des langues slaves et plus en général des langues non appartenant aux groupes roman et germanique, ou bien aux distinctions entre langues analytiques et synthétiques. Des remarques spécifiques ont intéressé également la langue française étudiée en Italie : dans les textes traduits dans cette langue, les étudiantes et les étudiants perçoivent qu'une indifférenciation est faite entre le féminin et le masculin dans les outils de TA, au détriment du féminin, ce qui est effectivement

[5] Rappelons que ces langues sont strictement liées aux parcours universitaires des étudiantes et des étudiants et aux langues qu'elles et ils apprennent.

le cas. À cet égard, ceux et celles qui étudient le droit (voir Bisiani) ont souligné la facilité majeure de l'outil de TA à traduire la langue anglaise, ainsi qu'une fiabilité majeure des résultats proposés lors de la traduction vers l'anglais. Par rapport à l'utilisation majoritaire de la langue anglaise vis-à-vis des autres langues dans l'industrie des langues, notamment dans le domaine de la TA et de la traduction plus généralement, la contribution de Vetere dans la première partie de cet ouvrage montre, d'ailleurs, que c'est cette langue à disposer de plus de ressources informatiques (et également de corpus alignés), d'où des performances meilleures et un meilleur entraînement des traducteurs automatiques et des résultats proposés pour cette langue.

Relativement aux outils terminologiques de l'IA connus et/ou utilisés, les questions posées ont visé avant tout les outils terminologiques utilisés en classe de langues. À propos des outils de gestion terminologique, quelques personnes déclarent connaître des programmes directement liés aux langues étrangères étudiées. En effet, lorsqu'on leur demande d'en préciser le nom, par une réponse non dirigée, sont cités ReversoContext, qui est surtout utilisé pour le français, Pons, qui est généralement utilisé pour la langue allemande, ou WordReference, qui est davantage utilisé pour la langue anglaise. Les réponses ont également montré une connaissance moyenne des outils terminologiques payants, comme ceux qui permettent l'alignement et l'identification de concordances (*Trados WinAlign*, *WordSmith Tools*, *Lingua MultiConcord*).

Au sujet des prédictions de saisie automatique lors de la rédaction d'un texte, d'un courriel ou d'un message sur les réseaux sociaux, les étudiantes et les étudiants sont plutôt douteuses et douteux. Si une partie de cette population tend à n'accepter la prédiction proposée que si elle la perçoit comme fiable, une partie presque égale préfère ne pas s'en servir. Plus rarement, les personnes ont déclaré effectuer des recherches supplémentaires avant de prendre une décision, ce qui devrait peut-être représenter le critère le plus logique à suivre pour bien choisir les mots à utiliser, sans *a priori* résultant d'une méfiance ou d'une indifférence à l'égard de la fiabilité de l'IA.

Le questionnaire proposé par de Vecchi a permis d'analyser un type particulier d'outil d'IA : le *chatbot* (*chat* d'assistance automatique), sorte d'« agent conversationnel »[6] qui est de plus en plus utilisé par les entre-

[6] Cf. « chatbot » dans le dictionnaire Larousse : https://www.larousse.fr/dictionnaires/francais/chatbot/188506

prises vendant leurs produits en ligne. Les participantes et les participants à l'enquête ont conscience du fait que ces dispositifs relèvent de l'IA, sans doute parce qu'ils sont utilisés en milieu professionnel et que plusieurs étudiantes et étudiants faisaient un stage en entreprise pendant l'administration du questionnaire. Les résultats des questionnaires montrent que les *chats* d'assistance automatique sont perçus aussi bien comme des instruments fiables que comme des outils de support à l'activité humaine.

Concernant l'automatisation des messages et des réponses, le questionnaire général a également permis de voir comment la population analysée perçoit l'intonation, la prononciation, la langue utilisée par les outils d'IA, et si l'on pouvait remarquer la présence majeure de voix masculines ou féminines dans les différents outils de réponse automatique. Les perceptions des étudiantes et des étudiants s'accordent sur l'inadéquation de l'intonation ou de la prononciation fournies par ces outils, parfois même sur la langue, considérée comme trop formelle ou inhabituelle, ainsi que sur l'emploi de voix masculines ou féminines. Ce dernier aspect est perçu comme le moins dérangeant. Par rapport aux *chats* d'assistance automatique, comme ceux des sites des opérateurs téléphoniques, c'est auprès des étudiantes et des étudiants en master qu'émerge l'utilisation assez fréquente de ces outils. Pourtant, elles et ils ne sont pas en mesure d'établir 100 % si les dialogues supposent la présence d'un interlocuteur humain ou de l'IA.

Quant à la question portant sur l'utilisation, dans les sites web, d'une langue inappropriée/inhabituelle, comme si elle était le résultat d'une traduction inappropriée ou erronée, les résultats montrent que les mots et expressions utilisés sont perçus comme s'ils étaient décontextualisés, autrement dit comme si leur traduction était « robotique », donc réalisée par la machine et, de ce fait, indifférente aux acceptions dont un mot peut se charger en contexte. D'autres ont évoqué une compétence insuffisante des traducteurs là où il n'y aurait aucune intervention humaine. Certaines personnes ont dénoncé la responsabilité des détenteurs des sites web, qui n'investissent pas assez pour pourvoir à la révision humaine de la traduction automatique. Pour ce qui est des éléments qui, selon cette population, sont traduits de la manière la moins correcte par les outils de l'IA, on trouve les jeux de mots, les expressions idiomatiques, les syntagmes nominaux complexes et les néologismes. Une fois de plus, le travail de l'hu-

main est donc perçu comme incontournable pour remédier aux erreurs provoquées par une traduction automatique non révisée.

Enfin, les réponses à la question relative aux aspects positifs ou négatifs caractérisant l'IA par rapport à l'apprentissage des langues — quoique variées dans leur formulation — mettent en évidence des éléments récurrents. Parmi les avantages, on a identifié la rapidité d'exécution de l'IA et la possibilité d'accéder aisément à des bases de données en ligne. Il est intéressant de remarquer que la population étudiant le droit en France identifie des avantages qui feraient des instruments issus de l'IA des outils « fédérateurs », permettant une meilleure communication entre les peuples, réduisant les barrières linguistiques et devenant ainsi un outil « antidiscriminatoire » et égalitaire. Curieusement, ces avantages ne sont présents que dans le premier questionnaire de ce groupe spécifique d'étudiantes et d'étudiants tandis que dans le second, ce même groupe a souligné la rapidité d'exécution des outils de TA, rejoignant ainsi les résultats préliminaires déjà identifiés par les étudiantes et les étudiants qui n'étudient pas le droit.

Parmi les désavantages, il émerge que le travail mené dans le cadre de la TA peut présenter deux types généraux de problèmes. D'une part, les désavantages concernent le manque de fiabilité des résultats des performances des dispositifs d'IA, mais aussi, en raison de la rapidité d'exécution de la traduction automatique, le manque de précision et de rigueur. D'autre part, l'aspect négatif le plus récurrent concerne la manière de repenser le travail de traduction humaine, avec les retombées que cela comporte en termes d'organisation du travail, de postes de travail perdus, voire la disparition même de la profession de traductrice ou traducteur, et cela au profit d'un produit final peu soigné mais moins coûteux pour les entreprises, les organismes ou les institutions qui s'en serviraient. D'où la nécessité de repenser également la formation et l'apprentissage des langues, puisqu'il ne serait plus utile ou rentable d'étudier une ou plusieurs langues étrangères si la communication passe quoi qu'il en soit, et parfois plus facilement, par des outils de TA.

Avant d'analyser les résultats des questionnaires de fin de cours, des réflexions sur ce qui émerge de la section relative aux langues minoritaires, à la manière dont on les perçoit et à leur emploi, dans la population étudiant le management en France (voir de Vecchi), sont incontournables. S'il est vrai que cet échantillon est unique et ne peut donc pas se prêter à

des comparaisons, il est quand même assez représentatif du fait du nombre potentiel de répondantes et de répondants, à savoir 609. Les questions qui ont été posées à cet égard concernent les langues régionales, les dialectes et/ou les patois parlés par les participantes et les participants. Les 403 réponses données ont été pour la grande majorité des cas négatifs (84,12 %), ce qui témoigne soit d'une méconnaissance à l'égard de ces langues soit que celles-ci sont en voie de disparition par marque de locutrices et de locuteurs ou bien par manque de prise en compte de la part des pouvoirs publics de leur possible extinction. Les réponses permettent également de constater une difficulté générale à définir ce qui est une langue minoritaire, comme leur statut « indéfini » en témoigne (voir la contribution d'Agresti dans la première partie de ce livre). Les cas, peu nombreux, où l'on affirme parler ces langues permettent de comprendre également ce qu'on considère comme étant une langue régionale, un dialecte ou un patois. On est généralement en mesure de reconnaître que, par exemple, des langues comme le picard, le créole, l'alsacien, l'occitan, ou les dialectes arabes — pour ne citer que les réponses les plus récurrentes — pourraient appartenir à cette catégorie, mais dès qu'on demande une définition de « langue minoritaire », les réponses décroissent, s'attestant à 230 sur le nombre total de 609. De plus, à l'égard de leur contenu, 15 % des répondants et des répondantes à cette question signalent ne pas savoir ce qu'est une langue minoritaire ou plutôt ne pas « s'aventurer » à essayer de la définir. Celles et ceux qui le font proposent le critère du nombre de locuteurs et de locutrices comme facteur qui peut définir la langue minoritaire ou définissent une langue minoritaire comme une langue « autre » par rapport à celle qui est perçue comme « dominante », à savoir la langue officielle du pays. Cette sensibilisation devient explicite dans la réponse déclarant qu'« aucune langue n'est minoritaire, elles sont toutes importantes » (voir de Vecchi), scellant ainsi que le critère permettant de classer les langues minoritaires est d'abord d'ordre qualitatif. Pour autant, lorsque cette population est invitée à donner un exemple de langue minoritaire, les réponses sont très variées, allant des langues régionales de France et des langues d'immigration — comprenant même des langues officielles, probablement parlées en famille et qui deviendraient par conséquent « autres » et « minorées » en France —, à des langues très éloignées de la France métropolitaine. Cependant, lorsqu'on affirme qu'une langue minoritaire est un patrimoine qui doit être protégé

par les institutions nationales et supranationales, on comprend que cette langue est perçue comme en danger.

Cette enquête préliminaire est suivie par des questions portant sur l'utilisation des langues minoritaires dans la vie quotidienne, dans les médias ou en relation avec l'IA. Presque la moitié des répondantes et des répondants à ce dernier aspect de la question souligne que les langues minoritaires pourraient tirer profit de l'IA, l'autre moitié répondant plutôt par « je ne sais pas ». Cela démontre, une fois de plus, que les langues minoritaires sont presque ignorées par la population interrogée. Celle-ci est amenée, par ses études ainsi que par les retombées immédiates en termes professionnels, à considérer la langue comme un instrument, et cela peut en partie expliquer les résultats des questionnaires. Enfin, selon la population interrogée par de Vecchi, ce sont les instances qui se focalisent sur les langues majoritaires à être responsable de l'abandon des langues « minoritaires ». D'ailleurs, les institutions éducatives devraient promouvoir cette richesse culturelle, comme quelques étudiantes et étudiants le proposent.

1.2 Résultats des seconds questionnaires

Pour ce qui est des questionnaires de fin de cours, il faut rappeler que, comme le montre le tableau 1, ils n'ont été proposés que dans le cadre de quelques cours de langue en Italie et dans celui de droit en France.

Nous signalons que les personnes qui ont rempli le premier questionnaire peuvent ne pas avoir rempli le second, comme le soulignent les chiffres du tableau 1. Le décalage de participation aux deux questionnaires pourrait être dû aux décrochages (abandon du cours universitaire).

Nous devons également rappeler que le second questionnaire pouvait contenir la section supplémentaire (Q2_AD) concernant le genre textuel (voir la sous-section homonyme de cette deuxième partie).

Par rapport aux réponses du premier questionnaire, il est possible de confirmer la majorité des pourcentages dans les réponses données précédemment, mais on trouve aussi des réponses différentes, ce qui témoignerait d'un changement de points de vue après avoir suivi le cours.

La plupart des remarques concernent la manière des dispositifs de performer lors de la TA. À l'égard de l'efficacité des solutions de traductions automatiques proposées, il émerge que c'est dans le domaine du lexique général que ces outils seraient le plus efficaces, tandis que l'inverse arriverait lors de la traduction du lexique spécialisé, des jeux de mots, des collo-

cations, jusqu'aux cas où les traducteurs automatiques seraient le moins efficaces, à savoir pour traduire la connotation des mots, les syntagmes nominaux complexes et les mots polysémiques. C'est là où les outils de TA sont insuffisants ou défaillants et que le travail de post-édition doit se focaliser davantage. À ce propos, les mots-clés les plus fréquents, qui émergent des réponses ouvertes que les étudiantes et les étudiants ont données, sont : « langue-culture », « idiomatique », « changement linguistique », « contexte », « créativité », « néologisme ». Bref, tout ce dont les outils de TA sont généralement dépourvus à l'heure actuelle.

Le travail de réflexion conduit pendant le cours a fait émerger les principaux défis auxquels les outils de TA et, plus généralement, l'IA se confronteront à l'avenir. Un aspect qui est strictement lié à ce dernier constat concerne la manière dont l'IA, notamment les outils de TA, agissent au niveau des genres textuels. Après avoir travaillé en cours sur certaines typologies textuelles (voir Cennamo, Mattioda ; Cinato ; Molino) et sur des langues différentes (le français, l'anglais, l'allemand), les étudiantes et les étudiants ont relevé que l'efficacité des outils de TA, qui est assez élevée dans le cas des textes à dominante informative et touristique, décroît lorsque le texte à traduire est de type juridique, pour atteindre des degrés d'efficacité encore plus bas dans les textes descriptifs et argumentatifs. Le type textuel le moins performant en TA serait le texte poétique, ces résultats étant en ligne avec les aspects linguistiques les plus simples/difficiles à traduire pour un outil de TA. De manière générale, plus la TA est utilisée pour traduire des mots isolés ou de courtes séquences de mots, plus la possibilité que la traduction proposée soit efficace augmente. Par conséquent, à la question (à réponse dirigée) de savoir si la TA est plus efficace au niveau de cotexte immédiat plutôt que sur le plan du contexte, la population interrogée est unanime : la TA est plus efficace, et peut-être plus fiable, au niveau du mot simple ou de séquence courte de mots plutôt qu'au niveau textuel, voire phrastique.

Conclusions et perspectives

Pour dresser un bilan, même provisoire, des résultats des questionnaires soumis aux étudiantes et aux étudiants qui ont participé à ce projet, et à la lumière de ce que nous venons de présenter, nous pouvons dire que les réponses au premier et surtout au second questionnaire sont liées

à l'utilisation en classe des outils d'IA, notamment de TA. À ce sujet, le travail en classe de langues sur ces outils et sur leur application à des domaines variés a enrichi les connaissances et les compétences des étudiantes et des étudiants et a contribué à développer leur sens critique par rapport à l'IA. Loin de nier l'utilité de la TA et des outils de l'IA au quotidien et dans tous les domaines, elles et ils comprennent que seul le travail effectué par l'humain est pour l'instant digne de confiance et de fiabilité, et qu'ils et elles sont en mesure de s'en apercevoir, identifiant comme « robotique » la traduction par la machine et comme « idiomatique » celle qui est réalisée ou bien révisée par l'humain.

Cette prise de conscience des outils de TA représente aussi la possibilité de renforcer la confiance dans ces outils, en améliorant leurs performances par l'activité de révision ou de post-édition finale.

Si l'intégration des outils de TA dans le domaine des enseignements universitaires, notamment de langue, représente désormais un impératif, c'est aux enseignantes et aux enseignants de langues de faire en sorte qu'on fasse une utilisation raisonnée et pertinente de ces ressources. La formation pourrait alors prévoir une systématisation de la post-édition à des fins d'apprentissage en s'appuyant sur la catégorisation de l'erreur en traduction (voir Cennamo, Mattioda). Cette dernière, en effet, permettrait à l'apprenante-traductrice ou à l'apprenant-traducteur de mieux connaître les outils de TA, en les examinant de façon critique. À cet égard, nous partageons l'avis de Massado et van der Meer (2017 : 15), qui remarquent que

> [l]'approche conventionnelle qui consiste à se demander quel genre de tâches un ordinateur ne sera jamais en mesure d'effectuer est un chemin risqué pour déterminer la façon dont l'être humain peut conserver sa valeur

et que c'est plutôt l'être humain qui doit être remis au centre pour ensuite réfléchir sur l'IA ou la repenser. Le travail de l'humain ne s'en trouve donc pas diminué, voire éliminé, comme quelques étudiantes ou étudiants le préconisent, mais il est à repenser en termes qualitatifs.

Le succès de participation à l'enquête menée par les questionnaires a permis de réfléchir à la poursuite du projet, notamment par la décision de soumettre de nouveau les deux questionnaires types (Q1_IT, Q2_FR) au premier et au second semestre de l'année universitaire 2021-2022. Cependant, par rapport à l'enquête ci-résumée, tous les questionnaires de base

qu'on soumettra à la fin du cours de l'an prochain contiendront des questions supplémentaires sur les aspects « genrés » qui peuvent concerner les outils informatiques utilisant des algorithmes d'IA. En outre, tous les questionnaires (Q1 et Q2 dans les deux langues) pourront prévoir des questions supplémentaires de plusieurs types (AD1, AD2, AD3…), concernant non seulement les genres textuels mais également d'autres aspects d'intérêt pédagogique et de recherche, comme, par exemple, la terminologie ou la transcription de l'oral à l'écrit. Les résultats de cette nouvelle édition de questionnaires feront l'objet de journées d'études de présentation des résultats et de publications futures, qui, nous espérons, auront le même ou encore plus de succès que l'initiative que nous avons essayé de présenter dans cette deuxième partie de l'ouvrage.

Bibliographie

Dörnyei Zoltán (2007). *Research methods in applied linguistics*. Oxford : Oxford University Press.

Massado Isabelle, van der Meer Jaap (2017). "Le marché de la traduction en 2022". *Rapport sectoriel du sommet TAUS*. Amsterdam : TAUS. Disponible à l'adresse https://mastertsmlille.files.wordpress.com/2018/06/the-translation-industry-in-2022_fr.pdf (dernière consultation le 29 novembre 2021).

Paveau Marie-Anne (2017). *L'analyse du discours numérique. Dictionnaire des formes et des pratiques*, Paris : Hermann, coll. « Cultures numériques ».

De Europa Deuxième partie - Special Issue 2022

Le multilinguisme européen et l'IA. Enquête auprès des futurs décideurs

Dardo de Vecchi

Introduction

Par rapport aux langues et dans l'enseignement supérieur, il est possible de distinguer deux groupes principaux d'étudiants : ceux pour qui les langues sont un matériau en soi (linguistique, langues, traduction, interprétation, littérature) et tous les autres, plus nombreux, pour qui les langues sont un moyen de dire, d'accéder aux connaissances ou plus tard de travailler, mais rarement un but en soi. La valeur des langues n'est pas la même pour les deux populations.

À l'intérieur du second groupe, diverses institutions les forment à occuper des postes de « décideurs » soit en entreprise soit, de manière plus vaste, dans des organisations, voire des administrations. Les étudiants de management ont des formations en finance, ressources humaines, marketing, commerce ou stratégie, etc. et notamment en l'international où les langues occupent *de facto* une place incontournable. Écoles de management, d'administration d'entreprise ou de gestion, dans une liste de dénominations non limitative, se retrouvent dans l'utilisation de ce qu'il est fréquent d'appeler « langue des affaires » censée en réalité englober une multitude de langues véhiculant des connaissances (langues spécialisées) dont l'approche est essentiellement pratique et bien différente de celle des étudiants du premier groupe.

Il en résulte que, dans les administrations et les entreprises, les décisions qui concernent, par exemple, la mise en place d'une politique linguistique ou de traduction en entreprise sont prises par des personnes dont les études ont été loin des préoccupations de celles des linguistes. C'est justement le cas des étudiants des écoles de management. De plus, les études de *International Business* (IB) ou management international connaissent un essor

Dardo de Vecchi, KEDGE Business School, dardomario.devecchi@kedgebs.com

depuis une vingtaine d'années, notamment depuis la parution en 1997 d'un article fondateur dont le nom était précurseur : « Language: The Forgotten Factor in Multinational Management ». Les études interculturelles se sont beaucoup développées, en même temps que la globalisation a mis en avant le besoin des langues. Elles sont approchées fréquemment dans une optique de communication et l'anglais *Business English Lingua Franca* (BELF) est vu comme obligatoire, incontournable et souvent comme une panacée.

Les habitants des pays où le multilinguisme est une réalité tangible sont habitués à intégrer cette coexistence des langues y compris pour des langues minoritaires. En revanche, dans d'autres pays, les langues vernaculaires et identitaires peuvent être voilées par une langue majoritaire ou des langues étrangères destinées aux seuls échanges internationaux. Le BELF voilerait ainsi l'existence de toute langue locale. Il est important de remarquer que les notions de langue minoritaire et de langue régionale ne sont pas très connues du grand public et de ce fait, elles sont souvent confondues. De plus, en France c'est la notion de langue régionale qui est plus connue. En même temps, la législation ne reconnaît pas de minorités, ce qui a comme conséquence que la notion de langue minoritaire est très peu perceptible, voire très peu comprise, comme nous le verrons plus bas.

Quoi qu'il en soit, la traduction prend une place importante en même temps que les technologies, dont l'intelligence artificielle (IA), se développent à grands pas. Toutefois, le BELF trouve des limites car pas toutes les populations le parlent. Les langues sont bien des ponts, mais également des frontières. Par exemple, lors de l'obtention de produits ou de matières premières où les connaissances de production sont exprimées en langue locale, et non en BELF.

Il nous semble alors important d'explorer l'image que ces managers futurs ont sur les sujets de la traduction, des langues minoritaires (LM) et de l'IA, car tôt ou tard ils devront décider s'il faut traduire, en quelles langues le faire, mais aussi réfléchir aux moyens dont ils disposent, avoir une politique linguistique ou encore le patrimoine linguistique qui doit être sauvegardé ou diffusé.

En fonction de l'expérience de 18 ans de cours en anglais et en français de *Management et langage* à la Kedge Business School, où, entre autres, les notions de dialecte, de variété et géographie linguistique sont abordées, nous avons pu affiner cette enquête qui fait écho aux connaissances que les étudiants ont partagées avec nous. Les technologies numériques occupent

une place importante dans les activités de formation de cette génération : rédaction des documents, même partagée, présentations des travaux, examens à distance, utilisation de plateformes pédagogiques, *serious games*, logiciels de téléconférence, etc.). Par ailleurs, l'utilisation des outils numériques concerne également les entreprises et les organisations où ces étudiants vont occuper des fonctions et où ils peuvent être amenés à se confronter à des langues minoritaires. Que ce soit dans l'enseignement ou dans l'entreprise, l'IA est de plus en plus présente et les étudiants, futurs décideurs, le savent bien. Si l'IA est utile pour les langues, tout comme elle contribue au travail quotidien des entreprises, il n'y a pas de raison pour penser que, confrontés à des langues minoritaires ou régionales, l'IA ne puisse pas proposer aussi des solutions efficaces. En effet, selon les pays, une langue vécue comme régionale, voire négligeable économiquement, en France peut avoir un statut différent dans un autre pays : c'est le cas du flamand dans le Nord, du basque dans le Sud-ouest ou du catalan au Sud. Ces trois langues ont des statuts bien différents de l'autre côté des frontières respectives. Là encore c'est notre expérience qui nous montre que ces différences sont balayées d'un revers de main : l'anglais sera utilisé en Belgique ou l'espagnol en Catalogne ou au Pays Basque. Là encore, si l'IA est utile pour le français, pour l'anglais ou pour l'espagnol, pourquoi ne le serait-elle pas pour ces autres langues ? Au pire, c'est l'anglais qui résoudra le « problème ».

Il nous a semblé important de mieux connaître, grâce à une enquête, les attitudes que ces managers futurs adoptent face à toutes ces thématiques ainsi que leurs représentations.

1. Public ciblé

L'enquête a été réalisée auprès des étudiants de cinq écoles de management en France, tous niveaux et programmes confondus. Cinq institutions, membres de la Conférence des Grandes écoles en France, ont participé : Kedge Business School, EDHEC, EM-Normandie, Montpellier Business School et l'École supérieure de commerce de Clermont. Les cinq établissements possèdent sensiblement les mêmes accréditations de référence reconnues internationalement dans l'enseignement du management : AACSB, EQUIS et AMBA[1].

[1] AACSB International—The Association to Advance Collegiate Schools of Business : https://www.aacsb.edu/. EQUIS, https://www.efmdglobal.org/accreditations/business-schools/equis/. Association of Masters in Business Administration (AMBA) : https://www.associationofmbas.com/

Réalisée en ligne avec le logiciel Qualtrics, l'enquête (voir Annexe) a été distribuée entre le 5 mars et le 21 avril 2021 et a reçu un total de 609 réponses.

2. Présentation de l'enquête

Un courriel proposant de participer à l'enquête présentait le cadre du projet européen dans lequel elle était réalisée : *Artificial Intelligence for European Integration* (AI4EI) et invitait à se connecter au logiciel *via* un lien hypertexte. Il a été possible de participer sur ordinateur, tablette ou téléphone afin de faciliter l'accès en toutes circonstances. Il a été fait mention que, en continuant, le participant acceptait de participer à l'enquête proposée.

61 questions ont été divisées en 5 secteurs : *Vous et l'IA, Vous et les langues, Vous et la traduction, Vous et les langues minoritaires* et *Histoire de vous connaître un peu*. L'articulation entre ces secteurs a suivi cet ordre afin d'obtenir un maximum de réponses nécessitant une réflexion plus poussée en début d'enquête. Le ton s'est voulu assez proche de l'oral pour éviter une barrière formelle qui empêcherait d'obtenir des données. Des énoncés ont été proposés pour que le répondant puisse être aiguillé dans ses choix, souvent multiples. Les options *oui, non, je ne sais pas* ont également été utilisées, mais en moindre mesure. Les demandes rédactionnelles ont été limitées, car elles exigent une réflexion de la part du répondant qu'il n'est pas en mesure de faire de manière immédiate ; cependant, des données intéressantes ont pu être collectées comme nous le verrons après.

3. Déroulement de l'enquête

3.1 Vous et l'intelligence artificielle

L'IA est un sujet devenu récurrent qui fascine autant qu'il est craint. Toutefois, son rôle devenant de plus en plus actuel et impactant, il est important de savoir la perception qu'en ont les jeunes générations.

Nous avons exploré les avis généraux à propos de l'IA à travers l'attitude vis-à-vis des chatbots[2] auxquels cette génération est habituée, la manière d'y écrire, la confiance, la perception du fait qu'il s'agit d'un

[2] La traduction en français de « *chatbot* » proposée par France Terme « dialogueur » est totalement inconnue en France auprès de cette génération et son utilisation aurait empêché l'obtention de données. Dans l'usage quotidien, les formes « chat », « bot » et « chatbot » alternent.

chatbot envisagé comme solution pour la communication des entreprises et si l'IA y joue un rôle majeur. Ensuite, nous avons cherché à connaître le rôle de l'IA pour l'amélioration de la traduction automatique, la traduction en général, l'enseignement des langues et les langues minoritaires. Toutes ces opinions, aussi brèves soient-elles, reposent sur une certaine vision de ce que sont les langues et que nous explorons dans le deuxième bloc de questions.

3.2 Vous et les langues

Pour cette population, dans la plupart des cas et en dehors d'un apprentissage précoce d'une autre langue ou de familles bi- ou trilingues, les langues étrangères sont apprises en cours et, comme pour beaucoup d'autres matières, leur apprentissage est sanctionné par une note. Or, dans la vie professionnelle, et pas seulement, les langues deviennent un élément du quotidien et cessent d'être un cours noté.

Nous avons recherché les raisons de l'utilisation d'autres langues en situation professionnelle (lire, rédiger, échanger oralement) ; la fréquence du ressenti de manques vis-à-vis de ces usages (à l'oral ou à l'écrit), de vocabulaire ou de confiance en soi. La représentation des langues a aussi été proposée : passion, nécessité, défi, contrainte d'apprentissage, moyen de travail ou approche culturelle. La mention de la langue maternelle apparaît en milieu de la section, puis la mention des autres langues apprises, y compris les langues régionales. Enfin, ce que ces langues régionales représentent pour le répondant : identité, enfance, famille, région, patrimoine, passé.

Il a été proposé de signaler en nombre de mois la durée des séjours à l'étranger dans la mesure où ils auraient pu contribuer à la confrontation à d'autres langues. L'enquête visait surtout à connaître les activités durant ces séjours : école secondaire, université, travail, *job* d'appoint, stage). En ce qui concerne les études et le travail, la difficulté d'utilisation d'une autre langue a été suggérée dans une échelle de Likert sur cinq points : très facile, facile, ni facile ni difficile, difficile et très difficile. Ensuite, un avis sur cette utilisation a été demandé (nécessité, défi ou autre avis), ce qui a paru particulièrement difficile : accents, vocabulaires nouveaux, vitesse d'élocution ou autres.

Il nous a paru intéressant également de savoir à quels sujets l'aide était demandée dans ces diverses situations (copain, collègue, *boss*, personne dans le même cas, quelqu'un de « sympa », application dans le *smartphone*,

site Internet ou autre). Nous avons demandé si, dans la rue, l'utilisation d'une autre langue était facile ou difficile en utilisant l'échelle précédente.

Situer ainsi les ressentis vis-à-vis des langues et de leurs utilisations permet de réaliser un pont vers le bloc suivant, qui concerne les traductions auxquelles les participants sont confrontés.

3.3 Vous et la traduction

La population que nous avons interrogée est sensibilisée à l'utilisation d'outils informatiques qui permettent de traduire. Elle n'est donc pas concernée par les problèmes théoriques de la traduction. Une certaine confiance serait accordée de fait aux outils utilisés.

Nous avons alors cherché les opinions concernant l'indice de fiabilité et de confiance de ces outils, si leurs résultats étaient vérifiés ou s'ils cherchaient à les vérifier, ou encore s'ils étaient perçus comme « bizarres ». Il nous a semblé prudent de chercher si des textes bilingues étaient attendus pour mieux rédiger des documents, si les sources des traductions proposées étaient vérifiées, si des dictionnaires bilingues en ligne ou spécialisés (par exemple de droit, finance, marketing ou ressources humaines) étaient consultés et s'ils connaissaient des bases de données terminologiques.

Finalement, il était important de savoir si, en situation professionnelle et devant un texte à traduire (il ne faut pas oublier que beaucoup d'étudiants sont en stage ou en apprentissage et des tâches leur sont confiées), ils traduisent eux-mêmes, car ils estiment pouvoir le faire, s'ils utilisent un traducteur automatique, s'ils confient le texte à un collègue ou s'ils font appel à un professionnel.

Dernière question concernant cette partie : le critère qui permet de choisir un équivalent dans une autre langue.

3.4 Vous et les langues minoritaires

Une fois les questions relatives aux langues et à la traduction vues, ce sont les langues minoritaires qui sont abordées. Nous avons demandé de dire ce qu'est une langue minoritaire et d'en donner des exemples. Il est important de remarquer qu'en France l'expression « langue minoritaire » est moins usitée que celle de « langue régionale » et que l'article de Wikipédia, source de consultation spontanée, titre : « Langues régionales ou

minoritaires de France »[3]. Une fois « situées », il est plus clair de proposer les items suivants à propos de ces langues[4] : elles doivent être protégées/ elles sont inutiles, doivent être enseignées à l'école/ abandonnées, considérées comme du patrimoine/ remplacées par d'autres, sont importantes pour le pays/ pour l'Union européenne.

Le lien à l'IA est fait en demandant si elle peut aider au développement des langues minoritaires. Ensuite, si ces langues régionales doivent être plus présentes dans les médias, si les entreprises devraient les mettre en valeur. Il est proposé de faire un commentaire à ce sujet. La dernière question propose de donner son sentiment : « Vous entendez quelqu'un parler en dialecte ou dans une langue régionale, vous vous dites : c'est folklorique, c'est rigolo, ça fait vulgaire, c'est intéressant, ça m'étonne toujours en bien, ça m'étonne toujours en mal ou autre impression ».

3.5 D'autres données statistiques : histoire de vous connaître un peu

Les critères classiques des enquêtes ont été recherchés. Vous êtes un homme, une femme ou « Je préfère ne pas répondre ». Les tranches d'âge ont été étalées entre moins de 18 ans, entre 18 et 25 (majorité des répondants) puis par tranche de 10 ans jusqu'à 85 ans et plus. Il ne faut pas oublier que le répondant à une enquête peut la distribuer à d'autres personnes. Les deux dernières mentions requises étaient le niveau d'études — années après le baccalauréat — ainsi que l'école où les études sont faites (donnée facultative).

4. Résultats de l'enquête

Pour rappel, nous avons globalement obtenu 609 réponses au questionnaire, mais certaines questions aucune réponse n'a été donnée. Pour cette raison, les graphiques présentent les quantités de réponses obtenues pour les items proposés. Lorsqu'un pourcentage apparaît, le nombre de réponses est signalé : Qn = Nombre de répondants (Ex. Q38 = 481) sauf dans les cas de réponses multiples.

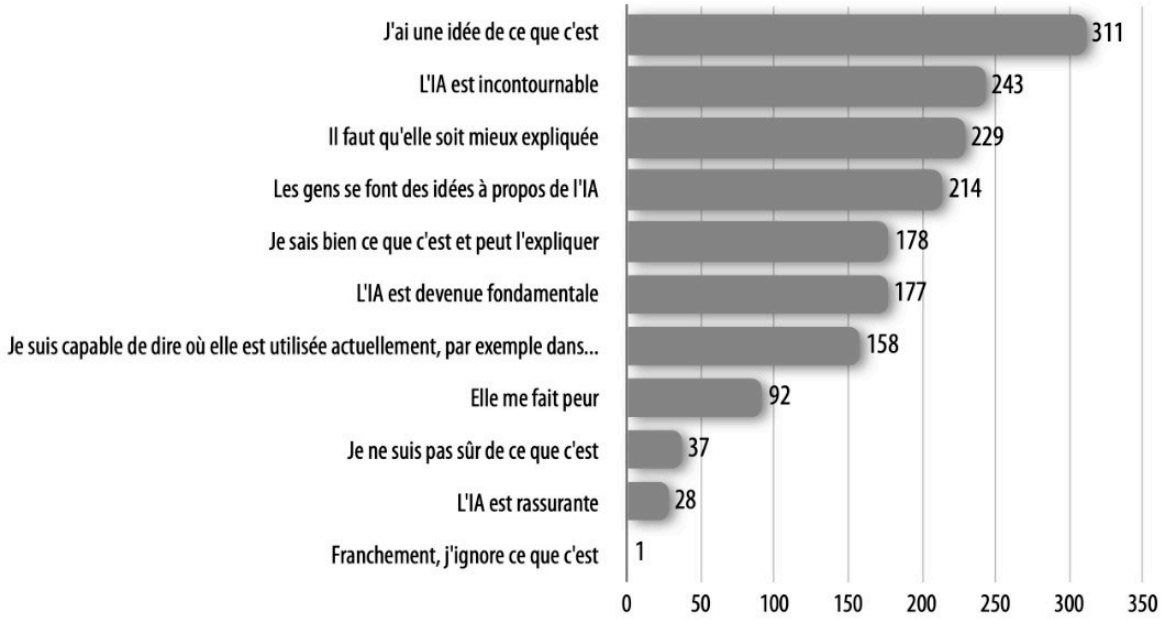
[3] https://fr.wikipedia.org/wiki/Langues_r%C3%A9gionales_ou_minoritaires_de_France, consulté le 2 mars 2021. À la même date, la simple recherche sur le moteur Google en français propose 975 000 résultats pour « langue minoritaire » contre 21 300 000 résultats pour « langue régionale ».

[4] Il est important de remarquer que notre enquête a été distribuée avant le débat en France à propos des langues régionales qui a eu lieu au mois de mai 2021.

4.1 Vous et l'IA

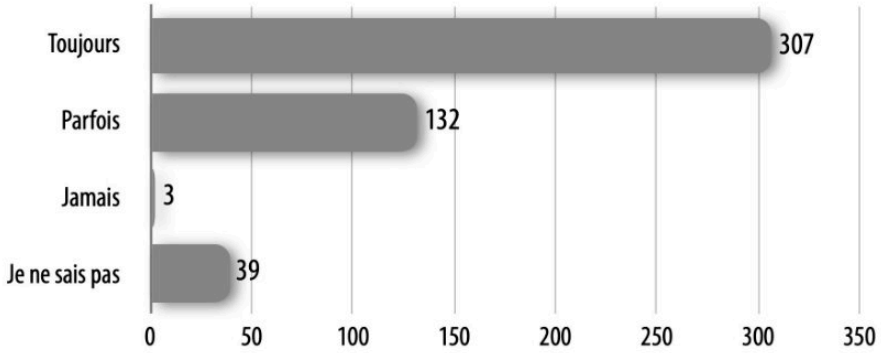
À propos de l'intelligence artificielle, un ensemble d'énoncés ont été proposés. Le graphique 1 montre les quantités de réponses pour chaque énoncé.

Les utilisations citées sont nombreuses, mais les plus fréquentes sont



Graphique 1 : Avis à propos de l'IA (Q32)

en ordre décroissant : la médecine, l'automobile, la finance, les téléphones, la reconnaissance faciale, les réseaux sociaux, le marketing, les chatbots, les moteurs de recherche, les traducteurs, les GAFAM[5].

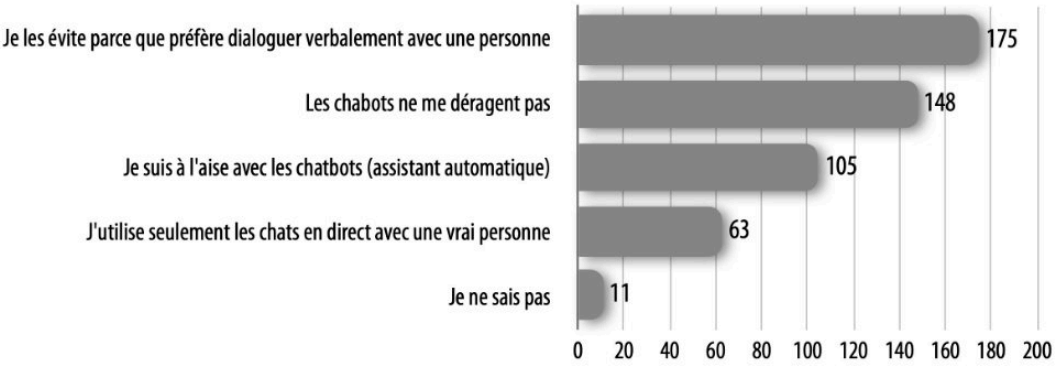


Graphique 2 : L'IA joue un rôle dans les chatbots (Q38 = 481)

[5] GAFAM : Google, Apple, Facebook, Amazon, Microsoft.

Cités par les répondants et fréquemment utilisés par les entreprises, les chatbots utilisent l'IA. Les répondants en sont conscients comme le montrent les scores à la proposition de la question 38 (graphique 2) : « L'IA joue un rôle majeur dans les chatbots ». Ils peuvent être utilisés comme « baromètre » de l'attitude envers l'IA.

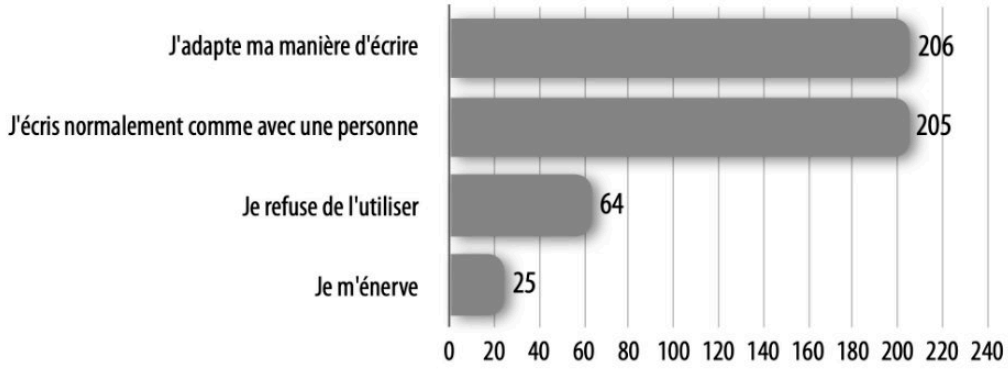
L'attitude envers les chatbots où l'IA est de mise apparaît dans les graphiques 3-6.



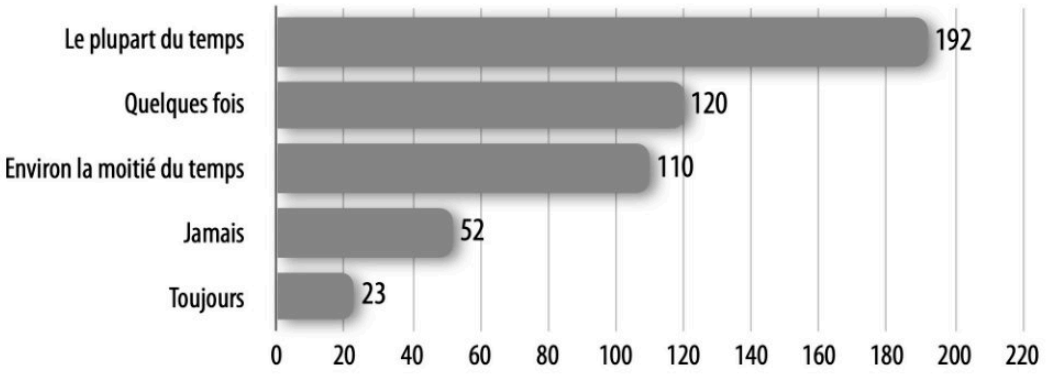
Graphique 3 : Lorsqu'une page web me propose un chat… (Q33 = 502)

Ces attitudes ne semblent toutefois pas conforter les scores des chatbots comme solution de communication, comme le montre le graphique 7.

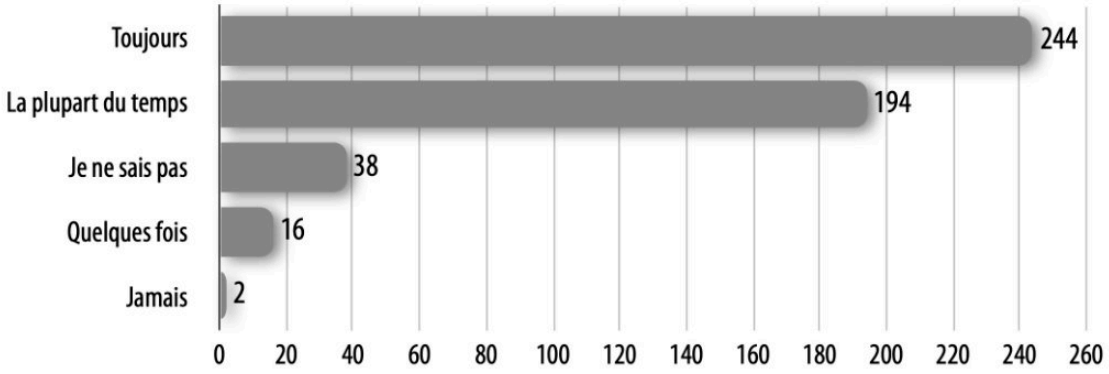
L'utilisation des chatbots comme application quotidienne de l'IA, notamment en milieu professionnel, permet de faire le lien à une autre application de l'IA par rapport aux langues : la traduction et l'enseignement, avant d'aborder le sujet des langues minoritaires.



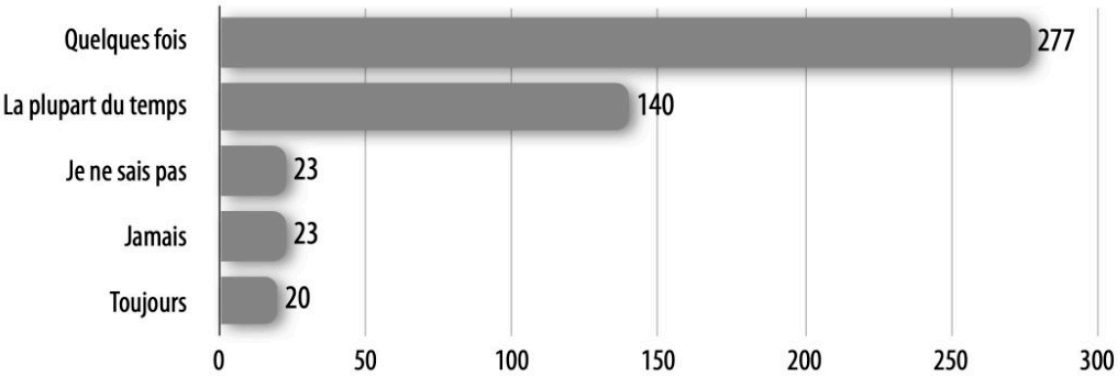
Graphique 4 : Lorsque j'ai à faire à un chatbot (Q34 = 500)



Graphique 5 : Je fais confiance à un chatbot (Q35 = 497)



Graphique 6 : Lorsque j'ai affaire à un chatbot, je m'en aperçois... (Q36 = 494)



Graphique 7 : Je pense que les chatbots sont une solution pour la communication des entreprises (Q37 = 483)

En lien avec la traduction automatique (Q39 = 469), 90,83 % pensent que l'IA peut améliorer la traduction automatique. 6,61 % ne savent pas et 2,56 % pensent que non. Ces chiffres sont proches pour le rôle que l'IA peut jouer dans la traduction en général (Q40 = 466) : 92,92 % de réponses positives, 1,93 % de réponses négatives et ne le savent pas : 5,15 %.

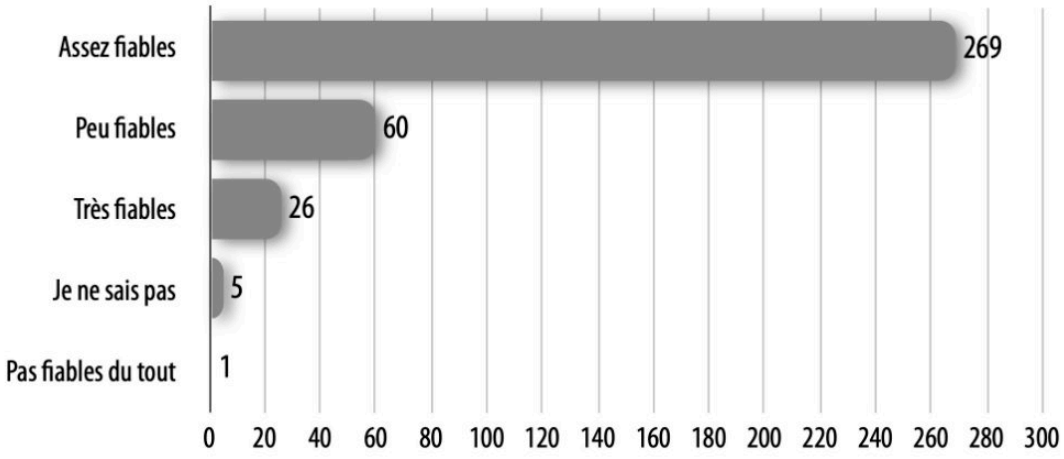
17,61 % pensent que l'IA ne peut pas jouer de rôle dans l'enseignement des langues contre 62,61 % qui pensent le contraire (Q41 = 460). 19,78 % ne le savent pas. Nous y reviendrons plus bas dans la section 4.3.

À la question de savoir si l'IA peut améliorer les applications d'enseignement des langues (Q46 = 429), la réponse positive atteint 80,19 %, la réponse négative 4,9 %, alors que 14,92 % ne le savent pas.

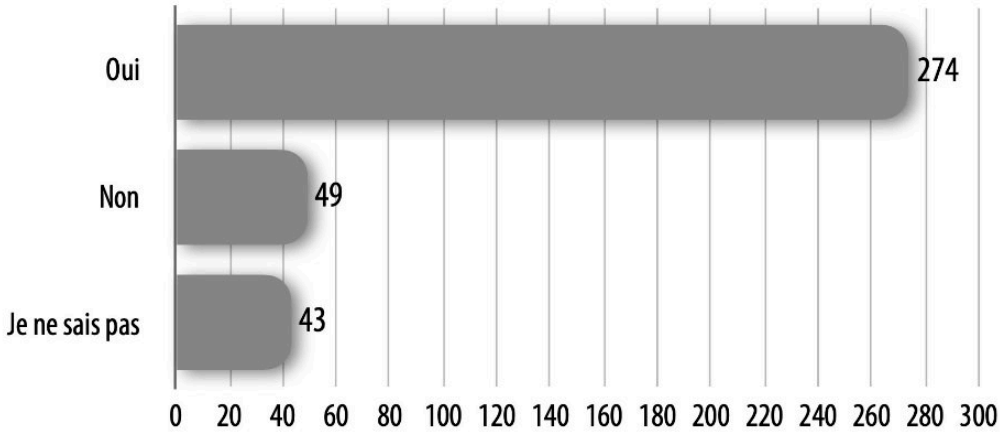
4.2 Vous et la traduction

En fonction de la vision à propos de l'IA, il est pertinent de savoir ce que cette population pense des outils de traduction, leur fiabilité et la confiance envers ces outils comme le montrent les graphiques 8 et 9.

Dans ce cadre, qu'en est-il des vérifications susceptibles d'être faites ? En ce qui concerne les résultats des outils automatiques (Q20 = 367), 55,31 % vérifient la plupart du temps, 29,70 % vérifient parfois et 14,99 %, rarement. 21,30 % trouvent la plupart du temps les traductions bizarres (Q21 = 366), 74,04 % parfois et rarement, 4,54 %. 51,91 % cherchent à vérifier ailleurs la plupart du temps, parfois 35,52 % et rarement 12,57 % (Q22 = 366).



Graphique 8 : La fiabilité des outils de traduction (Q18 = 361)



Graphique 9 : La confiance dans les outils de traduction (Q19 = 366)

Les données à propos de la recherche de textes bilingues pour améliorer les documents à rédiger (Q23 = 363) sont aussi intéressantes : 30,85 % le font la plupart du temps, 28,10 % parfois et 41,05 % rarement. Ces pourcentages sont à mettre en parallèle avec la vérification des sources : 21,79 % vérifient la plupart du temps, 38,55 % parfois et 39,66 % rarement (Q24 = 358).

Tenant compte de la population à laquelle nous nous adressons, il nous a paru important de savoir si les répondants connaissaient des dictionnaires bilingues en ligne (Q25 = 359). En effet, disposer de tels outils invite à mieux rédiger en langue étrangère. 73,82 % en connaissent, mais le reste (26,18 %) non. Curieusement, 19,50 % cherchent des dictionnaires spécialisés en ligne (Q26 = 359), 58,50 % n'en cherchent pas et 22,01 % n'en connaissent pas. En écho judicieux aux questions précédentes, seulement 4,76 % ont connaissance de bases de données terminologiques et ne savent pas de quoi il s'agit 52,94 % ; le reste (42,30 %) a répondu négativement (Q27 = 357).

8,78 % des enquêtés ont eu « énormément » d'expériences insatisfaisantes avec des traductions automatiques, 40,51 % beaucoup, 48,44 % un peu et seulement 2,27 % (Q28 = 353) jamais. D'ailleurs, seulement 12,10 % se souviennent d'un résultat marquant (en bien ou en mal). Le reste, pas du tout (Q29 = 347).

En situation professionnelle et devant un texte à traduire (Q30 = 347) : 70,89 % essayent de traduire eux-mêmes parce qu'ils estiment pouvoir le faire, 24,21 % utilisent un traducteur automatique, 4,32 % le

confient à un collègue qui connaît la langue et seulement 0,58 % font appel à un professionnel[6].

Lorsque les étudiants cherchent à traduire des mots en langue étrangère, 2,33 % choisissent le premier mot proposé, 54,94 % choisissent celui qui leur semble le plus approprié, 40,70 % pensent au contexte où il sera utilisé, les 2,03 % restants disent qu'ils se basent sur les contextes proposés ou sur la fréquence d'utilisation (cf. WordReference, Reverso) (Q31 = 344).

4.3 Vous et les langues

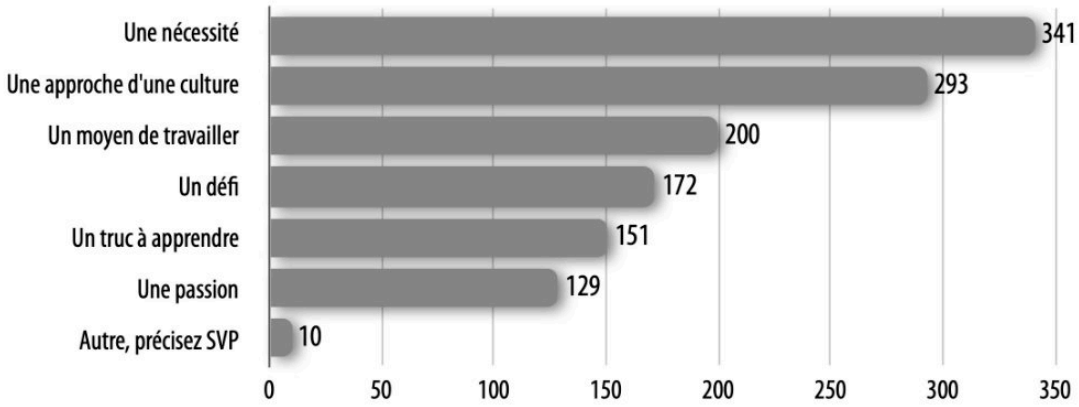
Il est important de connaître la place que les langues occupent dans une société donnée. Une société où le multilinguisme est fréquent n'a pas les mêmes réactions qu'une société essentiellement monolingue. Les réponses à nos questions, posées en français, ont été faites essentiellement par des étudiants français. Notre expérience d'enseignement en anglais nous montre les écarts d'opinion et d'attitude vis-à-vis des langues entre les étudiants venant des pays monolingues (France, Mexique, États-Unis, Brésil, Chine, etc.) et les étudiants originaires de pays multilingues ou présentant des variétés très marquées d'une même langue (Inde, Afrique du Sud, Maroc, Suisse). De plus, le regard porté sur les langues autres que la langue la plus diffusée dans le pays d'origine peut ne pas être le même.

Il nous a paru alors important de savoir ce que les langues représentent pour la population consultée, car cela pourrait avoir un rapport avec les langues minoritaires, ce qui apparaît dans le graphique 10.

Nous avons demandé quelle était la langue maternelle (Q3) et les langues secondes (Q4) afin de connaître l'environnement linguistique des répondants. Seulement 404 répondants de l'ensemble ont cité leur langue maternelle (le français dans 366 des cas, dans les autres 38 cas d'autres langues sont citées, dont 9 fois pour la plus fréquente[7]). Nous pouvons penser que

[6] Pour mieux comprendre ces chiffres, il faut se rappeler que les répondants sont souvent en stage ou en apprentissage et ils ne sont pas, par conséquent, en position de pouvoir décider de faire appel à un professionnel. En cours, il n'est pas rare d'avoir des étudiants, notamment parmi les apprentis, qui nous confient informellement que les entreprises leur confient des textes à traduire et lorsque ces étudiants sont bilingues, notamment du fait de leur situation familiale. Cela confirmerait l'hypothèse que les managers connaissent mal l'univers de la traduction et des langues.

[7] L'arabe cité 9 fois, le chinois 6, l'espagnol 6, le portugais 5, l'anglais 3 et l'italien 3. 6 autres langues ont été citées une seule fois.



Graphique 10 : Les langues sont… (plusieurs réponses possibles) (Q2)

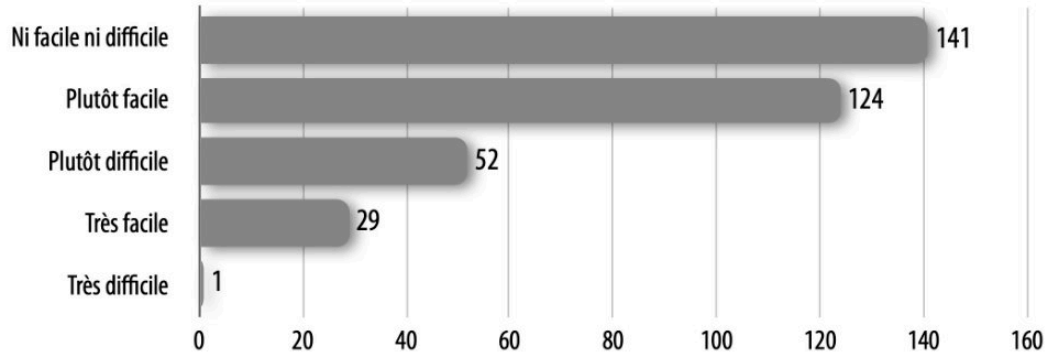
toutes ces langues ont été apprises en milieu familial et que, en conséquence, les autres sont des langues apprises à l'école et soumises à la sanction de la note : « j'avais de bonnes notes en… » étant un commentaire fréquent.

279 répondants déclarent connaître les langues secondes suivantes : anglais (185), français (31), espagnol (13), italien (9), arabe (7), allemand et portugais (6), chinois, cambodgien et créole (3). L'entrée « autres langues » a été citée une fois.

La pratique et la confiance dans les langues secondes peuvent se voir reflétées dans des séjours à l'étranger ou dans les diverses situations où l'utilisation de ces langues devient impérative. 93,98 % des personnes (Q8 = 399) ont réalisé un séjour à l'étranger durant les 5 dernières années dont la durée moyenne exprimée en mois était de 5,27 mois Q13). Les occasions de pratiquer une langue (Q10 : 354) à l'université (31,36 %), en stage (15,82 %), à l'école secondaire (11,58 %), au travail (5,65 %), dans un *job* d'appoint (2,26 %). Le reste de réponses (33,33 %) concerne essentiellement des séjours en famille, au pair, les vacances et le tourisme.

Les graphiques 11 et 12 montrent le degré de facilité/ difficulté à étudier et à travailler en langue étrangère, ce qui est important dans la mesure où la recherche d'outils informatiques peut y jouer un rôle.

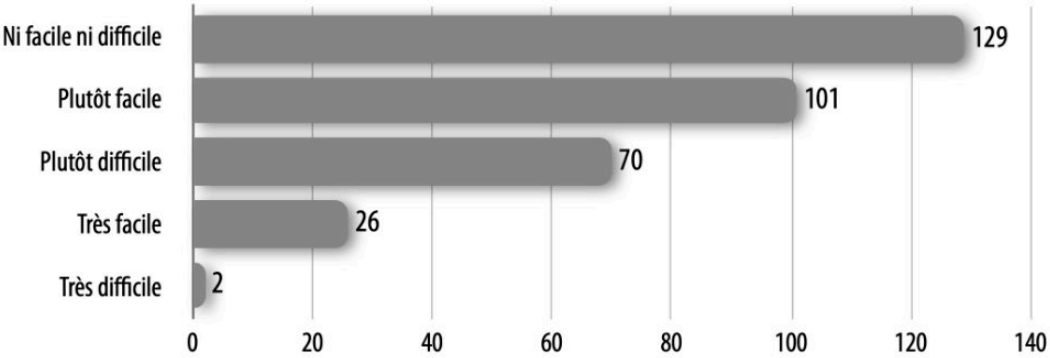
Les difficultés rencontrées (Q15 = 630, réponses multiples) étaient essentiellement l'accent des locuteurs locaux (30,16 %), les vocabulaires nouveaux (30,95 %), la vitesse d'élocution (35,56 %), le reste des réponses concerne dans l'ensemble : l'argot, des termes techniques, un alphabet différent, la nouveauté ou se voir répondre en français ou anglais lorsqu'on essaie de parler la langue (exemple donné, l'allemand).



Graphique 11 : Étudier en langue étrangère (Q11 = 347)

À la question : « Avez-vous déjà ressenti des manques au moment de vous exprimer à l'oral ? » 44,50 % avouent en ressentir (Q48 = 418). C'est le vocabulaire qui manque à l'oral ou à l'écrit[8] (59,76 %) (Q50 = 410), la confiance en soi 32,93 %, tandis que l'idiomaticité, la fluidité et la grammaire constituent l'essentiel des 7,32 % réponses restantes.

4.4 Vous et les langues minoritaires



Graphique 12 : Travailler en langue étrangère (Q12 = 328)

« Parlez-vous une langue régionale, un dialecte ou le patois ? » 403 personnes seulement ont répondu à la question[9] (Q6). Non : 84,12 %, oui, 4,22 % et parmi les langues citées (11,66 %) : picard (8), créole (7), alsa-

[8] Ces manques avaient déjà été rencontrés en 2005 à l'occasion d'une enquête auprès de la même population. Voir : de Vecchi (2008).

[9] Les options étaient : oui, non et si oui, laquelle ? Le pourcentage absent pouvant être dû au fait que nous n'avons pas voulu forcer les réponses pour passer à la question suivante, ce qui aurait pu avoir comme conséquence l'abandon de l'enquête.

cien (5), occitan (5), dialectes arabes (3). D'autres langues n'ont été citées qu'une seule fois (akyé, bavarois, bourbonnais, catalan, champenois, charentais, corse, franc-comtois, kabyle, laze, libanais, lingala, mahorais, swahili, tunisien, valencien et wengzhou).

La question 51 proposait : « Qu'est-ce que pour vous une 'langue minoritaire' ? ». Seules 230 réponses ont été obtenues. Parmi elles, 15 % de réponses franches : on ne sait pas de quoi il s'agit. 76,52 % s'accordent, avec des formulations très similaires les unes aux autres, à dire qu'il s'agit d'une langue « peu parlée ». Le reste des réponses (8,48 %) montre une variété de réponses dont : langue qui n'est pas commune à beaucoup de pays, non officielle, un dialecte local, langue dont la nécessité d'apprentissage est réduite ou peu parlée en dehors de son pays d'origine, peu connue dans le monde, parlée dans un seul pays, qui n'est pas principale dans un pays, langue d'un autre pays, non officielle, vernaculaire, non utilisée dans le domaine professionnel, langue d'une communauté minoritaire, moins utilisée ou en voie d'extinction. Une réponse attire notre attention : « aucune langue n'est minoritaire, elles sont toutes importantes ».

La question 52 proposait de donner des exemples de langues minoritaires. Les résultats obtenus peuvent être classés en quatre groupes : les réponses où une langue désignée est citée, les réponses générales, « je ne sais pas » et les réponses étonnantes. Les chiffres qui accompagnent correspondent au nombre de citations.

Pour le premier groupe, les langues suivantes ont été présentées comme minoritaires en ordre décroissant : breton (29), basque (21), corse (18), alsacien (15), suédois (10), catalan (9) finnois (9), grec (9), danois (8), flamand (7), polonais (7), russe (7), créole (6), picard (ch'ti) (6), italien (5), norvégien (5), néerlandais (4), occitan/ provençal (4), portugais (4), albanais (3), espagnol (3), gaélique (3), latin (3), lithuanien (3), roumain (3), aymara (2), bengali (2), biélorusse (2), cantonais (2), croate (2), hakka (2), hébreux (2), islandais (2), niçois (2), papou (2). N'ont été citées qu'une seule fois : allemand, auvergnat, bulgare, copte, galicien, gallois, ganyu, Hindou [*sic*], japonais, kikongo, kurde, liechtensteinois, lingala, maltais, mooré, napolitain, navajo, quechua, sarde, serbe, swahili, tagalog, turc, ukrainien, wallon, wolof, zoulou. Une réponse copiée-collée de Wikipédia n'a pas été prise en considération.

Les réponses générales, le second groupe : patois (25), dialectes (16), régionales (11) ; langues locales (5) ; langues scandinaves (2). Dans le troisième groupe, « je ne sais pas » nous avons eu 27 réponses.

Le dernier groupe est constitué de réponses pour le moins étonnantes qui sont révélatrices d'un manque de connaissance à propos des langues minoritaires : sont-elles des langues à part entière ? ; des langues « ethniques » [*sic*] ; langues de petits pays ; dialectes perdus ; dialectes de tribus d'Afrique (2 réponses) ; langues asiatiques (2 réponses) ; langues d'Afrique ; langues ancestrales ; ethnies chinoises ; langues amérindiennes (indiens d'Amérique) ; kabyle en France ; arabe au Royaume-Uni ; espagnol aux États-Unis ; français au Canada (6 réponses).

Lorsque nous avons cherché à connaître les opinions à propos des langues minoritaires, seulement 284 personnes ont répondu (Q53). Par ordre décroissant, les résultats sont les suivants, qui peuvent être séparés en avis positifs et négatifs.

Les avis positifs sont les suivants : une langue minoritaire doit être considérée comme patrimoine 46,83 %, protégée 31,34 %, est importante pour le pays 5,28 %, enseignée à l'école 2,82 % et importante pour l'Union européenne 2,11 % ; ces avis totalisent 88,83 % des réponses. Le nombre inférieur de répondants par rapport aux autres questions n'est pas sans rappeler qu'ils ne savent pas ce qu'est une langue « minoritaire ».

Les réponses négatives estiment qu'elles sont inutiles 2,46 %, doivent être abandonnées 1,76 % et remplacées par une autre langue 1,06 % totalisant 5,28 %. Les pourcentages restants concernent d'autres avis qui pouvaient ne pas se refléter dans la liste (6,34 %), mais des commentaires corroborent ce que nous avons entendu en cours. 7 « je ne sais pas » que nous pouvons interpréter de manière générale. Ensuite, les commentaires nous semblent intéressants : « a des risques [*sic*] de disparaître, mais cela ne veut pas dire qu'elle doit être protégée, elle disparaîtra naturellement si elle doit disparaître » ; « je ne sais pas ce que s'est [*sic*] » ; « je ne sais pas de quoi il s'agit, est-ce une langue en déperdition comme le français cajun ? » ; « Les langues suivent le cours du temps, elles s'adaptent » ; « Préserver son histoire mais l'abandonner si elle n'est plus parlée » ; « [elle doit être] pratiquée pour ne pas en perdre l'usage » ; « [elle doit être] protégée + considérée comme du patrimoine + importante pour le pays (surtout d'un point de vue culturel) [*sic*] ».

Concernant l'IA, 48,02 % pensent que l'IA peut aider le développement des langues minoritaires et 11,85 % pensent le contraire. 40,12 % ne savent pas (Q54 = 329).

Quant à la présence de ces langues dans les médias, 22,26 % des répondants pensent qu'elles devraient être plus présentes contre 30,27 % qui pensent le contraire et 47,48 % ne savent pas (Q55 = 337).

324 personnes ont répondu à la question 56 : pensez-vous que les entreprises devraient mettre en valeur les langues minoritaires ? Oui, pour 21,30 % et 39,20 % pensent le contraire, le reste (39,51 %) ne sait pas.

À la question 57 : « Avez-vous des commentaires à faire à propos des langues minoritaires ? » 106 commentaires ont été notés dans l'espace rédactionnel. Parmi eux, 56 commentaires se limitent à un simple « non » ou « je ne sais pas ce qu'est une langue minoritaire ». Les autres commentaires sont intéressants à plusieurs titres. Au premier chef, ils laissent apparaître que les notions de langue minoritaire, dialecte et patois ne sont pas correctement identifiées ou connues. Cela peut être influencé par le fait qu'en France, l'expression « langue minoritaire », et quelle qu'en soit la raison, n'est pas aussi diffusée que celle de « langue régionale ». En tout état de cause, les notions de patois et de dialecte ne sont pas vécues comme des formes de prestige. Nous pouvons identifier quatre catégories parmi les commentaires : positifs, négatifs, mitigés et autres.

Dans la catégorie des vingt commentaires positifs, apparaissent les idées suivantes : [elles sont un] point commun entre les personnes, [elles donnent un] sentiment d'appartenance et culture spéciale, il est regrettable de perdre certains accents, [elles sont] stigmatisées, besoin de pérennisation, il ne faut pas les perdre, [elles] servent à resserrer les liens dans un groupe de travail, les langues minoritaires représentent la richesse du passé, [elles] promeuvent l'identité d'un groupe, elles sont comparées à des œuvres comme dans les musées ou les bâtiments d'une ville, même si elles ne sont plus utilisées, elles sont importantes pour l'identité et l'authenticité d'un peuple, elles doivent être protégées, elles sont du patrimoine, [elles sont] importantes dans la mesure où l'on voudrait commercer avec le pays qui les parle, elles doivent être protégées. Un commentaire se démarque : « Toutes les langues peuvent être une langue minoritaire selon la situation géographique. L'inclusion est donc impérativement importante ».

Les quatre commentaires négatifs sont les suivants : « Elles doivent disparaître car elles peuvent se révéler dangereuses pour l'unité nationale ». « Ce n'est pas très important à mon sens ». « Créer une trop grande diversité au niveau des langues va créer des écarts de compréhension et

des exclusions dans certains domaines ». « Ça complexifie l'apprentissage d'une langue étrangère ».

Dans le troisième groupe, dix commentaires sont plus argumentés ; leur contenu étant particulièrement varié, il nous semble important de les transcrire en l'état. En effet, il serait dommage de les réduire aux idées principales.

« Doivent être conservées, mais il est inutile qu'elles prennent trop de place dans le monde professionnel. Chacun est libre d'apprendre la langue qu'il souhaite ». « Doivent être protégés parce que cela constitue le patrimoine d'une société, néanmoins on ne peut pas demander à tout le monde de s'y intéresser puisque ce n'est pas une langue internationale ». « D'un point de vue culturel, il est important de les conserver et observer la dynamique qu'elles offrent pour la culture qui les parle. D'un point de vue pratique, notamment la communication, elles posent beaucoup de difficultés et parfois même un frein pour les flux d'informations ou l'échange avec des personnes qui ne parle [*sic*] pas la langue (c'est pour ça que les langues minoritaires en entreprises ne devraient pas être mises en avant sauf si cela représente un enjeu stratégique) ». « Elles doivent faire partie de la culture générale et leur existence doit être diffusée et protégée. En revanche, il serait contreproductif de les utiliser en entreprise ou de manière générale au sein des institutions ». « Elles font partie de l'histoire, mais ne servent pas dans la vie quotidienne ». « Elles sont souvent réservées au cadre familial et sont menacées de disparition (pour autant, si ces langues minoritaires sont régionales, leur enseignement ne devrait pas être promu, au nom de l'union nationale) ». « Il ne serait pas judicieux de l'ajouter dans le business, peu de gens les parleraient et cela complexifierait encore plus les relations internationales. Par contre, d'un point de vue patrimoine, il peut être appris à l'école ou utiliser [*sic*] dans quelques journaux pour ceux qui souhaitent préserver la langue ». « Il y a suffisamment de langues comme cela sur Terre. Il est important de pouvoir traduire des langues peu utilisées voire mortes, et de pouvoir les comprendre, mais développer des langues minoritaires signifie très souvent sacrifier une autre langue (il est difficile d'apprendre deux langues en même temps). Par ailleurs, si toute la planète pouvait communiquer dans une même langue, avoir une langue secondaire commune, cela aiderait énormément. La priorité est, je pense, dans le développement d'une langue commune (l'anglais) et dans la réduction des langues minoritaires ». « Important à sau-

vegarder, mais pas forcément à imposer ». « Je dirai que c'est important qu'elles soient là, mais il faut parfois utiliser des langues classiques pour être sûr d'être compris ».

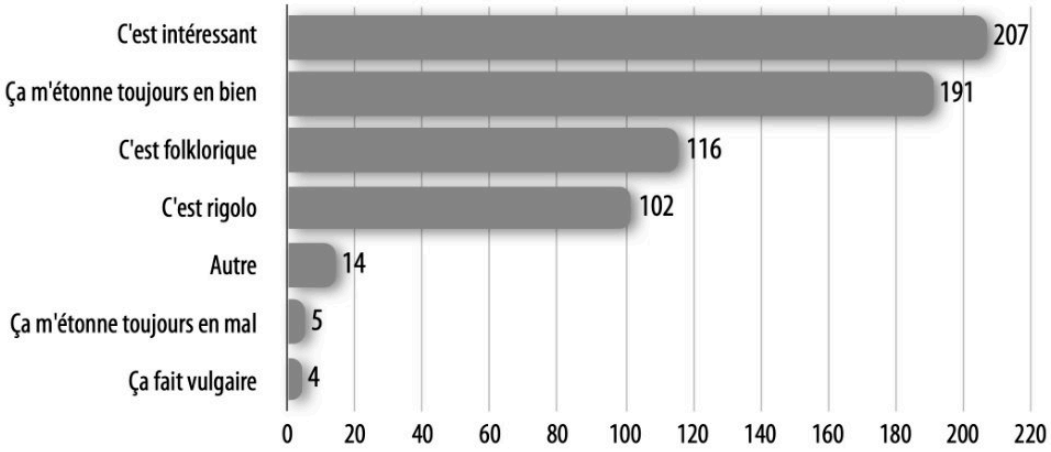
Le dernier groupe concerne sept avis divers. « Elles restent difficiles à développer, mais il ne faut pas tenter de les effacer ». « Il faudrait tout d'abord cerner le besoin de mettre en avant une langue minoritaire qui ne l'est qu'à l'extérieur de son environnement par définition. Est-ce le rôle d'une entreprise de le faire ? je ne pense pas. Cela est lié à la culture d'un pays ou d'une communauté ou de l'ethnologie. Je ne comprends pas le propos de ces dernières questions ». « Il s'agit plus de culture selon moi ». « Je pense qu'elles sont parfois négligées par les entreprises, probablement parce que même si on en maîtrise une, tellement peu de personnes la parle [*sic*] qu'on ne pourrait pas s'en servir. Mais je pense que cela reste un réel atout et un moyen de communication à ne pas négliger. Dans certains cas, rencontrer quelqu'un parlant cette même langue peut même créer de forts liens et rapprocher des personnes (= effet de communauté ?) [*sic*] ». « Patrimoine à sauvegarder, par le biais de la culture principalement, sans l'inclure dans les médias ou l'entreprise ». « Une fois que leur présence diminue, c'est très difficile de leur faire regagner de l'espace ».

La question 58 (graphique 13) était : « Vous entendez quelqu'un parler en dialecte ou dans une langue régionale, vous vous dites… » 639 réponses obtenues. Parmi les avis plutôt négatifs : ça fait vulgaire (0,63 %) et ça m'étonne toujours en mal (0,78 %). Aucun avis (2,19 %). Parmi les avis positifs (par ordre décroissant) : c'est intéressant 32,39 %, ça m'étonne toujours en bien 29,89 %, c'est folklorique 18,15 %, c'est rigolo 15,96 % totalisant 96,39 %. Il faut remarquer également que cette question a été traitée par la totalité des personnes sondées.

Sur 14 commentaires, 12 sont plutôt positifs et trois attirent notre attention : « Ça me paraît normal, vivant en pays catalan, c'est une habitude » ; « Parlant moi-même un dialecte, je ne pose aucune question. Chacun fait ce qu'il veut ! » et « Ça ne me fait rien, vu que je parle en dialecte marocain ». Ces commentaires nous invitent à penser que l'expérience du quotidien peut avoir une influence plutôt positive.

4.5 D'autres données statistiques : histoire de vous connaître un peu

339 personnes ont répondu à ces quatre dernières questions (Q59 à Q62), toutes ont souhaité répondre : femmes 60,77 % et hommes 39,23 %



Graphique 13 : Lorsque vous entendez quelqu'un parler en dialecte… (Q58)

(Je ne souhaite pas répondre 0 %). 91,20 % ont entre 18 et 24 ans, 7,04 % entre 25 et 34 ans et le reste 1,76 % entre 35 et 54 ans. Les plus grands pourcentages correspondent bien à la période d'études. Le niveau d'études en années après le baccalauréat français se répartit comme suit : +1, 10,82 %, +2, 11,40 %, +3, 20,47 %, +4, 24,27 %, +5, 31,58 %, au-delà, 1,46 %.

5. Discussion

Les outils informatiques se sont imposés dans la vie quotidienne, selon les âges et à des degrés divers. Toutefois, il est important de nuancer cette observation en fonction de leur nature et de l'intérêt que nous leur accordons. L'importance n'est pas la même pour l'utilisation d'un appareil électroménager ou pour les décisions prises par l'ordinateur de bord d'une voiture. De l'importance à la confiance, il n'y a qu'un pas.

Le chatbot est un outil informatique liant IA et facilité d'utilisation ; il assure la relation avec une entreprise ou une organisation. Il devient de plus en plus présent dans les sites Internet, applications, etc. et de ce fait constitue un bon élément immédiat de référence en relation avec l'IA. La tranche d'âge sollicitée dans cette enquête, celle des futurs décideurs, semblerait avoir des réticences quant à l'utilisation des chatbots, un outil de dialogue qui n'est pas sans rappeler le test de Turing. Nous avons vu les attitudes vis-à-vis des chatbots qui nous donnent des indices : les sondés les évitent (34,86 %), mais les réponses totalisent 50,40 % du panel. 40 % dialoguent normalement avec la machine, mais 41,20 % adaptent leur discours (conscience de l'existence du chatbot), bien loin devant ceux qui

refusent son utilisation (12,80 %). La confiance dans les chatbots (*toujours* et *la plupart du temps*) atteint 43,26 % mais serait seulement quelques fois une solution pour les entreprises à 57,35 % en dépit du fait que l'IA joue toujours un rôle majeur dans les chatbots (63,83 %). Confiance, d'accord, mais…

La place accordée à l'IA dans notre quotidien n'est pas aussi bien connue qu'on pourrait le croire (cf. « Il faut qu'elle soit mieux expliquée » : 13 % des réponses et « Les gens se font des idées à propos de l'IA », 12,83 %, des pourcentages proches). La presse écrite destinée au grand public ne s'empare que de temps à autre de la place qu'occupe l'intelligence artificielle dans nos vies. Marronnier pratique ou politique de l'autruche vis-à-vis de son réel impact ?

Il ne nous a pas semblé approprié de tenir compte des langues minoritaires sans approcher quelque peu l'image que les participants à l'enquête ont des langues de manière générale avant d'entrer dans ces cas précis. Par ailleurs, et dans le cadre d'études de plus en plus internationales, 93,98 % des personnes sondées pensent que les langues jouent un rôle important lors des séjours à l'étranger. On comprend alors pourquoi les langues sont une nécessité pour 26,37 % des personnes, ce qui n'est pas loin du pourcentage de ceux qui les conçoivent comme l'approche d'une autre culture (22,66 %). Les scores concernant la facilité pour étudier ou pour travailler dans une autre langue sont similaires, tout comme le fait qu'elles sont un défi ou une nécessité. Les difficultés concernent surtout le vocabulaire, les accents et la vitesse d'élocution.

Toutes ces situations (étude, travail, rue) confrontent cette population au besoin de traduire. Pour ce faire, les outils informatiques en lien avec l'IA deviennent incontournables et sont un deuxième point de rencontre entre la population interrogée et l'IA. Présents sur les *smartphones via* des applications ou en ligne, ils sont vécus comme fiables pour 74,52 % des répondants et la confiance en ces outils atteint 74,86 %. Néanmoins, ils vérifient la plupart du temps ET parfois à hauteur de 85,01 %, trouvant possiblement le résultat de la traduction bizarre (74,04 %), ce qui les inviterait à vérifier la plupart du temps ET souvent à hauteur de 87,43 %. La connaissance des outils tels que les dictionnaires bilingues et spécialisés existe, mais les bases de données sont visiblement très méconnues. Et pourtant, au travail, ils doivent traduire et ils le font eux-mêmes, car ils estiment en être capables à hauteur de 70,89 %. Cela invite à réfléchir sérieusement

sur la connaissance réelle que ces futurs décideurs ont du métier de traducteur et d'interprète ainsi qu'à la place que l'IA occupe véritablement dans les outils actuellement disponibles.

Dans nos enseignements, la méconnaissance des métiers liés à la traduction et à l'interprétariat n'a cessé de nous étonner depuis 2004, date à laquelle nous avions mis en place des cours de *Management et langage* au sein de la Kedge Business School. Notre enquête ne nous permet pas de savoir quelle est la raison de cette méconnaissance, mais elle mérite d'être recherchée sans tarder, car ce vide d'information nuit non seulement aux métiers concernés, mais aussi à la qualité des traductions dont beaucoup d'autres sujets dépendent.

D'après nos résultats, l'éternel tandem traditionnel langue/ culture semble se maintenir dans la culture française où le prestige de la langue occupe une place centrale qui relègue à des positions périphériques la présence des langues minoritaires, notamment régionales. Pour 84,12 % des sondés, en effet, elles ne sont pas parlées. Il serait difficile d'interpréter ce chiffre comme un rejet, car ces langues représentent toutefois des formes d'appartenance — ce qui n'est pas étonnant en soi — si on additionne l'identité, l'enfance, la famille et la région (66,67 %). Nous avons vu plus haut les images que les répondants ont des langues régionales, mais aussi leur méconnaissance de cet univers. Considérées cependant comme intéressantes, étonnantes et perçues comme du patrimoine, elles doivent être protégées, ce dont on ne peut que se réjouir, tout comme du fait que 48,02 % des répondants pensent que l'IA peut jouer un rôle les concernant, même si leur présence dans les entreprises — lieu de création lexicale — est mitigée (21,3 % de oui, 39,2 % de non, mais 39,51 % de « je ne sais pas »). Les langues régionales sont là, mais on ne saurait pas trop quoi en faire. Est-ce le poids de la langue de prestige ou la nécessité d'aller vers une autre langue (peut-être l'anglais) plutôt que de s'attarder dans une langue « peu parlée », citation fréquente ? Il ne nous est pas possible de trancher, mais nous avons pu voir qu'il y a une logique entre l'image de l'IA, l'attitude envers les langues, la traduction et finalement le rapport avec les langues minoritaires.

Conclusion

Une enquête, notamment en ligne, est toujours perfectible. Elle est dépendante d'un grand nombre de facteurs tels que le lieu et le moment où elle est réalisée, et à juste titre par cette génération hyperconnectée qui

peut y avoir répondu en toute circonstance. Néanmoins, les résultats obtenus grâce à ce panel spécifique confirment qu'un retour en arrière, ne serait-ce que vers une époque encore récente où l'outil informatique ne s'appuyait pas sur des algorithmes aussi développés et omniprésents, puis sur une intelligence artificielle incontournable, est impossible.

En matière de traductions, de langues en général et de patrimoine linguistique minoritaire ou régional en particulier, les futurs décideurs que nous formons aujourd'hui disposent-ils du recul intellectuel suffisant, voire de la maturité nécessaire, pour distinguer la valeur humaine de la langue, de sa transposition ou de son interprétation des apports technologiques de l'IA ? De plus, dans une accélération constante de traitement des données qui laisse peu de place à la réflexion (cf. Patino 2019) et où les premières propositions d'un moteur de recherche sont souvent acceptées sans être confrontées à d'autres et donc sans être analysées.

Dans l'introduction de *Se distraire à en mourir*, Neil Postman montre les deux visions du futur qui hantaient le vingtième siècle : celle d'Orwell avec *1984* et celle de Huxley avec *Le* meilleur *des mondes*. Il y écrit : « Orwell nous avertit du risque que nous courons d'être écrasés par une force oppressive externe, Huxley, dans sa vision, n'a nul besoin de faire intervenir un Big Brother pour expliquer que les gens seront dépossédés de leur autonomie, de leur maturité, de leur histoire. Il sait que les gens en viendront à aimer leur oppression, à adorer les technologies qui détruisent leur capacité de penser. » Tout laisse penser que nous y sommes déjà.

Bibliographie

Marschan Rebecca, Welch, Denice, Welch, Lawrence (1997). "Language: The Forgotten Factor in Multinational Management". *European Management Journal*, 15/5, 591-598.

Patino Bruno (2019). *La civilisation du poisson rouge*. Paris : Grasset.

Postman Neil (2011). *Se distraire à en mourir*. Paris : Fayard/Pluriel. (Trad. T. de Cherisey). Version originale : Postman Neil (1986). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. New York : Penguin.

(de) Vecchi Dardo (2008). « Place de la terminologie dans l'enseignement des langues spécialisées dans une école de management ». Dans : Olivier Bertrand, Isabelle Schaffner (dir.). *Le français de spécialité. Enjeux culturels et linguistiques*. Paris : Éditions de l'École Polytechnique, 237-249.

Annexe : détail de l'enquête Intelligence artificielle, langues, traduction et langues minoritaires

NB : les questions sont présentées ci-dessous dans l'ordre auquel le répondant les a lues. Le numéro correspond à l'ordre des questions dans le logiciel Qualtrics. Certaines questions n'ont pas été exploitées, car les réponses n'étaient pas pertinentes pour l'analyse ci-dessus.

Bonjour,
Merci d'avoir accepté de participer à cette enquête qui nous permettra de mieux connaître vos avis sur les langues, l'intelligence artificielle et la traduction. Cette enquête est réalisée dans le cadre du projet européen *Artificial Intelligence for European Integration* (AI4EI). L'enquête est anonyme et vous prendra environ 10 minutes. En continuant, vous acceptez de participer.
Encore une fois, merci !
Dardo de Vecchi, Kedge Business School

32 L'intelligence artificielle. Cochez les phrases avec lesquelles vous êtes d'accord.
Je sais bien ce que c'est et je peux l'expliquer
J'ai une idée de ce que c'est
Je ne suis pas sûr de ce que c'est
Franchement, j'ignore ce que c'est
Elle me fait peur
Il faut qu'elle soit mieux expliquée
Les gens se font des idées à propos de l'IA
L'IA est devenue fondamentale
L'IA est incontournable
L'IA est rassurante
Je suis capable de dire où elle est utilisée actuellement, par exemple dans…

33 Lorsqu'une page web me propose un chat :
Je suis à l'aise avec les chatbots (assistant automatique)
Je les évite parce que préfère dialoguer verbalement avec une personne
J'utilise seulement les chats en direct avec une vraie personne
Les chabots ne me dérangent pas
Je ne sais pas

34 Lorsque j'ai à faire à un chatbot,
J'écris normalement comme avec une personne

J'adapte ma manière d'écrire
Je m'énerve
Je refuse de l'utiliser

35 Je fais confiance à un chatbot :
Toujours, La plupart du temps, Environ la moitié du temps, Quelques fois, Jamais

36 Lorsque j'ai affaire à un chatbot, je m'en aperçois…
Toujours, La plupart du temps, Environ la moitié du temps, Quelques fois, Jamais

37 Je pense que les chatbots sont une solution pour la communication des entreprises
Toujours, La plupart du temps, Environ la moitié du temps, Quelques fois, Jamais

38 L'IA joue un rôle majeur dans les chatbots.
Toujours, Parfois, Jamais, Je ne sais pas

39 Pensez-vous que l'IA peut améliorer la traduction automatique ?
Oui, Non, Je ne sais pas

40 D'après vous, l'IA peut-elle jouer un rôle dans la traduction en général ?
Oui, Non, Je ne sais pas

41 D'après vous, l'IA peut-elle jouer un rôle dans l'enseignement des langues ?
Oui, Non, Je ne sais pas

42 Avez-vous déjà utilisé une application d'enseignement de langues ?
Oui, Si oui, laquelle ? Non, Je ne sais pas

43 À propos de ces applications et globalement diriez-vous que vous êtes…
Convaincu, Assez satisfait, Déçu, Je ne sais pas

44 D'après vous, ces applications vous permettent de :
Pratiquer, entendre, lire, avancer, avoir confiance en moi, Je ne sais pas

45 Qu'est-ce qui vous manque dans ces applications ?
Des explications grammaticales
La possibilité de poser des questions
L'acquisition de vocabulaire
La possibilité de vérifier ma prononciation

Autre (précisez SVP)

46 Pensez-vous que l'IA peut améliorer ces applications ?
Oui, Non, Je ne sais pas

47 En situation professionnelle, lorsque vous devez parler une autre langue, il s'agit essentiellement pour :
Rédiger des textes, lire des textes, parler en direct avec des collègues, clients, etc., autres (précisez SVP)

48 Avez-vous déjà ressenti des manques au moment de vous exprimer à l'oral ?
Souvent, Parfois, Rarement

49 Avez-vous déjà ressenti des manques au moment de vous exprimer à l'écrit ?
Souvent, Parfois, Rarement

50 À l'oral ou à l'écrit ce qui vous manque c'est :
Le vocabulaire, de la confiance en vous, Autre (précisez SVP)

52 Pour vous les langues sont…
Une passion, une nécessité, un défi, un truc à apprendre, un moyen de travailler, une approche d'une culture, Autre, précisez SVP

****

3 Ma langue maternelle est le…

4 Ma seconde langue est le… (Ex. avec mes parents je parle en français, mais je vivais en Allemagne et parlais aussi l'allemand)

5 Ensuite, j'ai appris le… (Ex. études, religion, etc.)

6 Parlez-vous une langue régionale, dialecte ou patois ? (Ex. enfant, je vivais en Alsace et je comprends l'alsacien).
Oui, Non, Si oui, laquelle ?

7 Vous parlez une langue régionale, un dialecte ou un patois. Pour vous, celareprésente…

Mon identité, Mon enfance, Ma famille, Ma région, Mon patrimoine, Le passé, Autre (précisez SVP)

8 Avez-vous fait des séjours à l'étranger durant les 5 dernières années ?

Oui, Non

9 Votre séjour à l'étranger a duré…
[Échelle allant de 0 à 24 moins proposée avec un curseur]

10 Pendant mon séjour à l'étranger,
J'ai été à l'école secondaire, j'ai été à l'université, j'ai travaillé, j'ai eu un job d'appoint, j'ai fait un stage, Autres cas (précisez SVP)

11 Étudier dans une autre langue a été…
Très facile, plutôt facile, ni facile ni difficile, plutôt difficile, très difficile

12 Travailler dans une autre langue a été…
Très facile, plutôt facile, ni facile ni difficile, plutôt difficile, très difficile

13 Étudier dans une langue étrangère était pour moi…
Une nécessité, un défi, pas d'avis

14 Travailler dans une langue étrangère était pour moi…
Une nécessité, un défi, pas d'avis

15 Ce qui m'a paru difficile c'était…
L'accent des locaux, les vocabulaires nouveaux, les gens qui parlent vite, Autre (précisez SVP)

16 Pour mieux comprendre, je posais des questions à…
Un copain, un collègue, un professeur, mon « boss », une personne dans mon cas, quelqu'un de « sympa », une application dans mon smartphone, un site Internet, Autre (précisez SVP)

17 Parler dans la rue c'était plutôt…
Très facile, plutôt facile, ni facile ni difficile, plutôt difficile, très difficile

18 Je pense que les outils de traduction en ligne ou les applications sont
Très fiables, Assez fiables, Peu fiables, Pas du tout fiables, Je ne sais pas

19 Les outils de traduction en ligne vous inspirent-ils confiance ?
Oui, Non, Je ne sais pas

20 Vérifiez-vous le résultat des outils automatiques ?
La plupart du temps, Parfois, Rarement

21 Il vous arrive de trouver qu'une traduction automatique est « *bizarre* » ?
La plupart du temps, Parfois, Rarement

22 Cherchez-vous à vérifier ailleurs ?
La plupart du temps, Parfois, Rarement

23 Cherchez-vous des exemples de textes bilingues, pour mieux rédiger vos textes ?
La plupart du temps, Parfois, Rarement

24 Lorsqu'on vous propose des exemples, vérifiez-vous les sources ?
La plupart du temps, Parfois, Rarement

25 Connaissez-vous des dictionnaires bilingues en ligne ?
Oui, Non, Je n'en connais pas

26 Cherchez-vous des dictionnaires spécialisés en ligne ? (Ex. droit, finance, marketing, ressources humaines, etc.)
Oui, Non, Je n'en connais pas

27 Connaissez-vous des bases de données terminologiques ?
Oui, Non, Je ne sais pas de quoi il s'agit

28 Avez-vous eu des expériences de traduction automatique insatisfaisantes ?
Énormément, Beaucoup, Un peu, Jamais

29 Vous souvenez-vous d'un exemple en particulier (bon ou mauvais) ?
Oui, celui-ci : …, Non, je ne me souviens pas

30 En situation professionnelle et devant un texte à traduire :
Vous essayez de traduire vous-même parce que vous estimez pouvoir le faire, Vous utilisez un traducteur automatique, Vous le confiez à un collègue qui connaît la langue, Vous faites appel à un professionnel

31 Lorsque vous utilisez un système de traduction en ligne pour chercher un mot, selon quel principe choisissez-vous l'équivalent dans la langue étrangère ?
Je choisis le premier terme proposé. Je choisis celui qui me semble le plus adapté. Je pense au contexte où j'utiliserai le terme. Autre (précisez SVP)

51 Qu'est-ce que pour vous une « langue minoritaire » ?

52 Pouvez-vous donner des exemples de langues minoritaires ?

53 Pour vous, une langue minoritaire :
Doit être protégée, est inutile, doit être enseignée à l'école, doit être aban-

donnée être considérée comme du patrimoine, être remplacée par une autre, est importante pour votre pays, est importante pour l'Union européenne, autre (précisez SVP)

54 Est-ce que pour vous l'IA peut aider le développement des langues minoritaires ?
Oui, Non, Je ne sais pas

55 Est-ce que les langues minoritaires doivent être plus présentes dans les médias :
Oui, Non, Je ne sais pas

56 Pensez-vous que les entreprises devraient mettre en valeur les langues minoritaires ?
Oui, Non, Je ne sais pas

57 Avez-vous des commentaires à faire à propos des langues minoritaires ?

58 Vous entendez quelqu'un parler en dialecte ou dans une langue régionale, vous vous dites :
C'est folklorique, c'est rigolo, ça fait vulgaire, c'est intéressant, ça m'étonne toujours en bien, ça m'étonne toujours en mal / Autre

59 Vous êtes… un homme / une femme / ne souhaite pas répondre

60 Votre âge…
Moins de 18 ans, Entre 18 et 24 ans, Entre 25 et 34 ans, Entre 35 et 44 ans, Entre 45 et 54 ans, Entre 55 et 64 ans, Entre 65 et 74 ans, Entre 75 et 84 ans, 85 ans ou plus

61 Vous êtes à Bac+… 1, 2, 3, 4, 5, +

62 Quel est le nom de votre établissement (facultatif) ?

De Europa Deuxième partie - Special Issue 2022

Les dispositifs de traduction automatique et la recherche terminologique comme outils pédagogiques pour des étudiant·e·s en droit

Francesca Bisiani

Introduction

Quelle est l'interaction entre la terminologie multilingue, le droit et les dispositifs de traduction automatique ? Voici le questionnement principal qui a nourri notre projet didactique mené de janvier à avril 2021 à la Faculté Libre de Droit (FLD) de l'Université Catholique de Lille.

Dans le cadre de l'espace multilingue européen, et plus généralement dans les sociétés plurilingues, l'harmonisation des concepts soulève plusieurs problématiques liées aux dynamiques différentes qui régissent chaque système juridique national[1]. L'intégration européenne passe donc, entre autres, par une interdépendance entre les langues et le droit qui représente depuis des années un enjeu majeur pour les services de traduction et de terminologie de l'Union européenne. Or, toute cette production terminologique émise par les institutions ne se répand pas seulement dans des discours variés (médiatique, juridique, politique) sur le plan national et international, mais aussi dans des dispositifs de traduction automatique neuronale qui puisent dans des corpus multilingues existants pour entraîner leurs algorithmes d'apprentissage.

Le projet que nous avons proposé à la FLD, et que nous allons présenter dans cet article, a été conçu à partir de cette réflexion interdisciplinaire qui puise donc à la fois dans les sciences juridiques et linguistiques. L'articulation de ces champs d'études nous a permis de poursuivre un double objectif pédagogique tout au long de notre étude. Le premier, de type empirique, s'inscrit dans les intérêts de recherche de l'axe linguistique intitu-

Francesca Bisiani, Université Catholique de Lille, francesca.bisiani@univ-catholille.fr

[1] Pour un aperçu historique de l'interaction entre les langues et le droit et pour des exemples de recherche dans ce champ, voir le numéro intitulé « Langue et droit : terminologie et traduction » de la *Revue française de linguistique appliquée,* coordonnée par Philippe Gréciano et John Humbley (2011).

lé « *Linguistic rights and language varieties in Europe in the age of artificial intelligence* » du Projet européen *Artificial intelligence for European integration* (AI4EI)[2]. Il vise à la fois à mener une enquête, par l'administration de deux questionnaires auprès des étudiant·e·s sur leur usage des outils de traduction automatique, et à fournir des éléments pour maîtriser consciemment ces dispositifs dans leur champ d'études. Le deuxième objectif, de type théorique, cherche à initier les étudiant·e·s en droit au raisonnement terminologique par un travail individuel de recherche qui veut conduire, comme but ultime, à la création d'une base de données de termes juridiques[3]. L'hypothèse sur laquelle s'appuie notre travail est que la confrontation entre les recherches qualitatives des étudiant·e·s et les résultats de la traduction automatique — ainsi que, plus largement, l'observation d'un concept juridique en plusieurs langues — puissent sensibiliser le public universitaire à la diversité linguistique et terminologique.

Dans la première partie de cet article, nous présentons le cadre général de notre étude pédagogique et l'approche théorique qui a été utilisée lors de l'expérimentation. Nous passerons ensuite, dans la deuxième partie, à l'observation de deux exemples d'études terminologiques effectuées par les étudiant·e·s et aux résultats de l'enquête par questionnaire. Dans les conclusions, nous reviendrons sur notre hypothèse de départ, ce qui nous permettra de formuler quelques considérations sur l'intérêt de travailler sur la variante terminologique avec un public d'étudiant·e·s en droit.

1. Présentation du projet didactique à la Faculté Libre de Droit

L'étude didactique que nous avons menée a été réalisée avec 38 étudiant·e·s inscrit·e·s au Master 1 « Organisations internationales et européennes » et au Master 2 « Droits de l'Homme, Sécurité et Développement » de janvier à avril 2021. Ces deux Masters bilingues français/ anglais font partie de l'*International and European Law School* de la Faculté Libre de Droit à l'Université Catholique de Lille et sont axés sur le fonctionnement du système juridique international.

[2] Le Projet AI4EI est dirigé par Umberto Morelli de l'Université de Turin et financé par la Commission européenne (2020-2023). L'axe linguistique est coordonné par Rachele Raus de l'Université de Bologne.

[3] Le projet en cours de création de la base de données s'inscrit dans le cadre des activités du CR3D, le Centre de recherche sur les relations entre le risque et le droit, de la FLD de l'Université Catholique de Lille.

Il importe donc de souligner que les étudiant·e·s qui ont participé au projet s'intéressent tout particulièrement au droit international, au droit européen et aux droits humains et c'est justement dans ces domaines d'étude qu'il·elle·s ont puisé pour sélectionner les termes objet de leur étude. Rappelons que, selon les principes théoriques de la terminologie, « les *termes* sont des *unités lexicales* dont le sens est envisagé par rapport à un *domaine de spécialité* »[4] (L'Homme 2004 : 22). Le point de départ de notre projet a été le choix des termes, qui a été effectué en amont de la recherche et qui, dans la plupart des cas, a été fait à partir des sujets des mémoires des étudiant·e·s[5] ou en lien avec les thématiques proposées dans les cours magistraux des parcours concernés. À la suite de cette première étape, les participant·e·s ont procédé à la rédaction des fiches terminologiques et à l'analyse comparative des résultats affichés par les traducteurs automatiques selon des critères déterminés, que nous allons présenter dans le paragraphe suivant.

Les deux questionnaires communs au groupe de recherche du Projet AI4EI[6], qui nous ont permis d'effectuer une enquête sur l'usage et l'opinion des étudiant·e·s à l'égard des dispositifs de traduction automatique, ont été présentés au tout début et à la fin du cours. L'analyse comparative des données nous a permis de retracer l'évolution de la réflexion des participant·e·s et, en même temps, de mettre en avant certaines problématiques liées à la traduction automatique des concepts juridiques. Nous expliquons d'abord les bases théoriques de notre travail pour passer ensuite à la présentation de deux recherches terminologiques.

2. L'approche théorique et la rédaction des fiches terminologiques

Le cadre théorique que nous avons mobilisé tout au long de notre expérimentation relève de la terminologie, et plus particulièrement des démarches sémasiologiques de ce domaine d'études. La théorie générale de la terminologie du XX[e] siècle prend son essor dans les années 1950-1960 (Zanola 2018 : 26), notamment à partir des travaux d'Eugène Wüster[7]

[4] En italique dans le texte original.

[5] Tel est le cas des recherches que nous allons présenter au paragraphe 3.

[6] Nous nous référons ici au groupe de recherche sur la didactique de l'axe linguistique intitulé « *Linguistic rights and language varieties in Europe in the age of artificial intelligence* ».

[7] Le travail pionnier de E. Wüster, considéré comme fondateur de la discipline, est sa thèse soutenue en 1931, dont le titre en français est « La normalisation linguistique internationale en technologie, en particulier en électrotechnique ».

dans une époque caractérisée par le développement des échanges internationaux dans le domaine de l'industrie et de la technique et par la recherche d'une communication sans ambiguïté (Zanola 2018 : 27).

Dans l'approche proposée par Wüster, appelée également classique ou onomasiologique, le travail terminologique doit viser à la désambiguïsation qui caractérise les mots de la langue afin de limiter la variation ou la synonymie et, par conséquent, d'assurer une communication plus efficace, notamment dans les systèmes industriels et de production. En d'autres termes, bien que Wüster observe les variations et les « éléments d'instabilité » (Candel 2004 : 20) de la langue, il encourage la normalisation et la biunivocité de la terminologie à des fins de catégorisation des savoirs techniques. En effet, il faut considérer que, durant cette époque pionnière de la terminologie, les travaux de Wüster sont axés sur l'harmonisation de la terminologie de type industriel et considèrent donc en moindre mesure les variations textuelles de la terminologie juridique. Par la suite, à partir des années 1960, nous assistons, sous l'impulsion des avancées technologiques en matière d'extraction terminologique et de linguistique du corpus, à un bouleversement de la théorie classique (Humbley 2018a : 80). À partir de cette époque, les détracteurs de cette dernière remettent la variation, et donc la pluralité des dénominations qui peuvent être conférées au même concept, au centre du débat. Sans entrer dans le détail de cette démarche, appelée sémasiologique, il convient de souligner que la lignée de travaux qui se développent à partir de ces principes se tourne progressivement vers la dimension contextuelle et discursive du terme.

Notre projet didactique s'aligne sur ces théorisations et s'intéresse tout particulièrement aux contextes d'usages du terme et, dans une optique discursive, aux remaniements du terme selon les positionnements discursifs des énonciateur·trice·s (Maingueneau 2002 : 453)[8]. Il s'agit concrètement d'observer la présence (ou l'absence) du terme recherché dans plusieurs contextes (juridique, politique, médiatique) et d'examiner les variantes dénominatives qui ressortent du travail de recherche terminographique.

[8] Le positionnement est une notion qui relève de l'analyse du discours de l'École française (Dufour, Rosier 2012). Parmi les approches discursives de la terminologie, la démarche d'« archive » de Rachele Raus (2013) s'intéresse tout particulièrement aux positionnements des énonciateur·trice·s qu'on peut retracer à l'aide des variantes terminologiques.

Pour ce faire, les étudiant·e·s ont été invité·e·s à remplir des fiches en plusieurs langues, sous forme de tableau, qui serviront, à moyen terme, à construire une base de données terminologique[9]. Ces fiches contiennent plusieurs champs qui se répartissent selon les quatre catégories principales proposées par Juan C. Sager (1990) et reprises par Marie-Claude L'Homme :

- ***Données conceptuelles***[10] : On regroupe ici la définition, l'indication de relations avec d'autres concepts, le domaine de spécialité et, au besoin, des notes techniques et des illustrations, etc. [...]
- ***Données linguistiques*** : Ce second groupe comprend la ou les formes linguistiques proprement dites (termes, synonymes, variantes, etc.), l'information grammaticale qui s'y rattache, et, au besoin, des marques d'usage.
- ***Données pragmatiques*** : Il s'agit ici des contextes servant à illustrer l'emploi des termes, des indications de langue, ou toute autre mention relative aux conditions d'utilisation des termes.
- ***Données relatives aux équivalents*** : Les données linguistiques rattachées aux équivalents dans une autre langue font partie de cette dernière catégorie (L'Homme 2004 : 254).

Parmi toutes ces informations, nous souhaitons mettre en avant l'importance que nous avons accordée aux données pragmatiques dans les fiches terminologiques.

Les étudiant·e·s ont été appelé·e·s à signaler des types de contextes qui avaient été prédéfinis et intitulés en anglais de la manière suivante : *international legal context, european legal context*, *national legal context*, *mass media context*, *political context 1*, *political coxtext 2 (counter argument)*. Pour chaque champ, les participant·e·s ont donc dû procéder à une recherche documentaire afin de trouver et copier dans la fiche de travail un contexte contenant le terme choisi, tout en indiquant la typologie de texte concerné (par exemple, texte normatif contraignant, non contraignant, jurisprudence). Ce type d'exercice, effectué en deux ou trois langues selon les connaissances linguistiques des étudiant·e·s, nous a permis de faire res-

[9] Nous précisons que le projet de la construction de la base de données se trouve actuellement dans sa phase initiale qui consiste à collecter les données par la création de fiches terminologiques. L'encodage des premières données dans un système de gestion informatique est prévu pour l'automne 2021.

[10] L'italique et le gras apparaissent dans le texte original.

sortir toute une série de phénomènes langagiers et discursifs qui ont été évalués à la fin du travail par l'analyse comparative des données collectées, comme nous le verrons un peu plus loin dans le présent paragraphe. À titre d'exemple, l'observation d'une dénomination donnée dans plusieurs sources juridiques et sur plusieurs plans peut faire émerger des désalignements interprétatifs, intralinguistiques ou interlinguistiques, suggérer l'absence de normalisation d'un concept juridique international ou européen dans le système juridique national ou encore montrer la présence de variations dénominatives. En même temps, l'analyse des contextes politiques peut révéler des débats qui existent autour d'un concept donné, ce qui peut se manifester, par exemple, par des opérations de reformulation[11] ou de « silenciement » (Pulcinelli-Orlandi 1996 : 62) d'une dénomination.

Pour revenir sur la structure de la fiche de travail, aux quatre catégories principales, nous avons ajouté trois autres rubriques intitulées : *machine translation*, *linguistic analysis*, *legal analysis*. Dans la première partie, dédiée à la traduction automatique (TA), les participant·e·s ont procédé à l'analyse et à la comparaison des résultats affichés sur certains dispositifs de TA. Plus précisément, il·elle·s ont recherché le terme faisant l'objet de la fiche bilingue ou trilingue dans les traducteurs automatiques Google Translate et/ ou DeepL, mais aussi dans les concordanciers Linguee et/ ou Reverso Context. Ces enquêtes nous ont permis de comparer les résultats obtenus pendant la rédaction de la première partie de la fiche multilingue (les données conceptuelles, linguistiques, pragmatiques et relatives aux équivalents) avec les équivalents proposés par les dispositifs. Dans le cadre des objectifs du projet didactique, cette démarche expérimentale s'est révélée utile, car elle a encouragé les étudiant·e·s à prendre conscience de la variation terminologique, notamment des effets performatifs liés au choix de la dénomination dans le domaine juridique. Par exemple, un étudiant en Master 1, H.T., a constaté que le terme anglais « *whistleblower* » sur Linguee et Reverso Context est principalement traduit par l'équivalent à connotation négative « dénonciateur»[12]. Or, cette variante n'est pas toujours appropriée dans le droit interne français, car elle est susceptible de véhiculer une connotation

[11] Par exemple, une étudiante, V-M. L., a observé que dans certains cas le terme « système d'armes létales autonome » était reformulé par les détracteurs de ces armes par la variante « robot tueur ».

[12] La dernière vérification sur le site de Linguee et de Reverso Context a été effectuée le 13/07/2021.

négative et de ne pas prendre en compte la raison qui pousse la personne intéressée à divulguer des informations concernant des actes illicites, à savoir l'intérêt général. Dans certains contextes, la variante la plus adaptée serait donc « lanceur d'alerte », qui renvoie à une catégorie protégée en France par la loi, dite « Sapin 2 », sur la transparence, la lutte contre la corruption et la modernisation de la vie économique.

D'un point de vue théorique, il est intéressant de remarquer que dans cette rubrique intitulée « *machine translation* », notre approche terminologique s'entrecroise avec une pratique réflexive qui relève de la traductologie et de la communication multilingue. Il s'agit, en ce sens, d'analyser, dans une perspective cognitive, « les produits de la traduction [...] à la lumière des rapports de force idéologiques » (Guidère 2008 : 89) et en relation aux systèmes culturels et juridiques d'arrivée. Dans cette optique, la traduction automatique est considérée comme un relayeur du discours et de la terminologie proposée par les logiciels de traduction qui puisent dans des bases de données existantes, surtout institutionnelles, pour entraîner les algorithmes intelligents. Les reprises traductives des dispositifs ne sont donc pas neutres, car elles amplifient des discours ou des choix traductifs qui ont été effectués précédemment par des énonciateur·trice·s humain·e·s.

Les deux autres parties, la *linguistic analysis* et la *legal analysis*, sont consacrées aux remarques des étudiant·e·s et aux réflexions effectuées pendant le travail de recherche terminologique. Ces analyses argumentées visent à faire ressortir les difficultés rencontrées (par exemple au moment de trouver une définition), les observations concernant les variations (surtout leur lien avec les positionnements politiques des énonciateur·trice·s) ou les débats qui entourent le concept qui fait l'objet de la fiche sur le plan international, européen ou national. Ces sections ont également été utiles pour aborder la question des valeurs différentes (par exemple, argumentative, perlocutoire) des dénominations selon le type de texte dans lequel le terme est mobilisé. L'analyse des équivalents ou de la présence ou absence d'un terme donné sur le plan intralinguistique permet tout particulièrement de réfléchir sur la valeur jussive que véhicule la terminologie juridique dans les textes normatifs ou dans la jurisprudence à effet contraignant. Cette valeur traduit « l'intention de l'émetteur de faire accomplir au destinataire un acte déterminé » (Pascual 2004 : 34) et invite à s'interroger sur les choix dénominatifs des énonciateur·trice·s en relation avec les spécificités du texte et du contexte analysés.

Avant de passer à la présentation de deux travaux, qui nous serviront d'exemple pour illustrer la démarche suivie dans notre projet didactique, nous souhaitons clarifier un dernier point concernant notre approche théorique. Bien que notre optique s'appuie sur des critères sémasiologiques, la démarche onomasiologique n'est pas exclue, notamment dans le cas des néologismes. En effet, pour certaines fiches contenant des néologismes, les étudiant·e·s ont parfois été amené·e·s à proposer des dénominations ou des définitions pour des concepts qui sont déjà connus en anglais, mais qui n'ont pas encore d'équivalents stables en français ou dans les autres langues cibles. Cela relève d'une démarche qui puise à la fois dans l'onomasiologie et la sémasiologie[13] et qui a pour objectif de suggérer l'harmonisation d'un concept, tout en faisant attention au contexte d'usage.

Au paragraphe suivant, nous exposons brièvement les points saillants de deux recherches effectuées par deux étudiantes du groupe en Master 2 à partir des dénominations « *feminicidio* », en espagnol, et « minorité », en français. Notre attention se focalisera tout particulièrement sur la rubrique des fiches terminologiques intitulée « *machine translation* » et donc sur les résultats obtenus par les dispositifs de traduction automatique.

3. Deux termes controversés : « *feminicidio* » et « minorité »

Le premier travail que nous allons présenter concerne le terme « *feminicidio* » et a été réalisé par V.A.C.G. en anglais et en espagnol.

L'analyse de V.A.C.G.[14] met tout d'abord en avant la controverse qui entoure la définition et la dénomination du concept, notamment par rapport au terme « *femicidio* ». Tel que l'explique Marylène Lapalus (2015 : 86), « les concepts de feminicidio et femicidio sont aujourd'hui couramment utilisés sur le continent latino-américain » bien que leur définition et leur promotion « demeure[nt] un enjeu fondamental dans la visibilisation de ces violences ». La conceptualisation des deux termes s'élabore à partir de plusieurs mouvements militants féministes qui sont strictement liés aux différents contextes historico-culturels. « *Femicidio* », dans son acceptation contem-

[13] Sur l'onomasiologie comme méthodologie pour étudier la néonymie, voir Humbley (2018b).

[14] Les deux travaux ont été présentés au *Workshop* du 24 avril 2021 organisé dans le cadre du colloque « Droit et variétés linguistiques en Europe à l'aune de l'intelligence artificielle » du Projet AI4EI. Nous avons anonymisé par des sigles les noms des deux étudiant.e.s qui y ont participé.

poraine, dérive du terme « *femicide* », forgé par les féministes nord-américaines Jill Radfort et Diana Russel, qui définissent en 1992 ce concept comme le meurtre misogyne commis par un homme envers une femme (Radford, Russel 1992). En revanche, le terme « *feminicidio* » a été proposé par Marcela Lagarde (2008)[15] dans le cadre des mobilisations au Mexique contre les assassinats de plusieurs femmes à Ciudad Juárez au début des années 2000. Ce qui distingue « *feminicidio* » de « *femicidio* » selon l'auteure, c'est surtout l'impunité des auteurs des crimes et la « violence institutionnelle » (Lagarde 2006 : 221) et systémique qui caractérise les meurtres de femmes au Mexique. Ainsi, M. Lagarde élargit le concept de « femicide » et attribue l'une des causes de ces meurtres aux défaillances des autorités publiques (Zanotta Machado 2019) et à la non-protection des victimes par l'État.

Si ces termes ont été largement empruntés et réutilisés dans d'autres contextes politiques et juridiques, nous constatons que les deux dénominations ne peuvent pas être considérées sémantiquement comme équivalentes, d'autant plus que ces termes tendent à accompagner des discours militants qui sont liés à des contextes spécifiques.

À partir de cet arrière-plan, l'étudiante a pu remarquer que ces concepts sont parfois utilisés de façon interchangeable, notamment dans le contexte international et par certains traducteurs automatiques. Elle a dressé un tableau terminologique, dont nous avons tiré ces quelques extraits dans le tableau 1 :

International legal context (Spanish)	"[...] El análisis de los datos debería permitir la identificación de errores en la protección y servir para mejorar y seguir desarrollando medidas de prevención, [...] para la recopilación de datos administrativos sobre los asesinatos de mujeres por razón de género, también conocidos como "femicidio" o "feminicidio", y los intentos de asesinato de mujeres."	Source : Naciones Unidas (2017). Recomendación general num. 35 sobre la violencia por razón de género contra la mujer, por la que se actualiza la recomendación general num. 19. https://www.acnur.org/fileadmin/Documentos/BDL/2017/11405.pdf. (Consulté le 17/02/2021)
Machine translation (without context)	1. feminicidio – femicide (ES-EN Google Translate) 2. feminicidio – femicide (EN US) DeepL 3. feminicidio – feminicide (ES UK) DeepL	Sources : 1. Google Translate 2. DeepL 3. DeepL
Machine translation (with context)	4. México informó de un programa concreto dedicado al feminicidio. 4. Mexico reported a specific programme focused on femicide. 5. Varios estados han introducido la figura del feminicidio en sus códigos penales. 5. Several states have introduced the category of feminicide into their penal codes. 6. Junto con este discurso conceptual, las feministas mexicanas decidieron recurrir directamente al origen latino del término y llamarlo "feminicidio". 6. Parallel to this conceptual discourse, Mexican feminists decided to translate the term femicide directly from its Latin origins as "feminicidio".	Source : Reverso Context

Tableau 1 : Extrait de la fiche terminologique de V.A.C.G. sur le terme « feminicidio »

[15] M. Lagarde traduit l'ouvrage de J. Radfort et D. Russel de 1992 et c'est justement à cette occasion-là qu'elle propose, avec l'accord des auteures, le terme « *feminicidio* » pour souligner l'adaptation du concept au contexte mexicain et pour le distinguer des concepts de « femicide » ou d'« homicide d'une femme » (Lagarde 2006 : 221).

Dans les rubriques dédiées à la *machine translation*, la dénomination en espagnol « *feminicidio* » est parfois traduite en anglais par « *femicide* » (ex. 1, 2 et 4) et d'autres fois par « *feminicide* » (ex. 3 et 5). Le traducteur automatique ne semble pas toujours prendre en compte le système juridico-politique d'arrivée ou les catégories notionnelles permettant de distinguer les deux termes, comme la question de l'impunité ou de la misogynie qui caractérisent cet acte meurtrier. Il est néanmoins intéressant de noter qu'au point 6 du tableau, le concordancier Reverso Context ne traduit pas le terme « *feminicidio* », qui est mis entre guillemets, et permet, par conséquent, de révéler l'ancrage du terme dans le contexte militant mexicain.

Le deuxième travail, qui a été effectué par G.T., se penche sur le concept de « minorité » en français et de « *minority* » en anglais. Dans ce travail, l'étudiante s'intéresse au débat qui entoure la représentation des minorités (par exemple, linguistiques, ethniques, sexuelles ou autre) dans le droit international et dans le droit français et à la traduction automatique de ce terme. Il faut d'abord rappeler que ce concept n'a pas de définition précise au niveau du droit international public, ce qui est principalement dû « à la réticence des gouvernements à accepter que le droit international régi[t] une affaire considérée comme interne » (Abou Ramadan 2016 : 79) et aux difficultés d'englober dans une même définition une pluralité de situations. Cela dit, les institutions internationales et européennes utilisent cette dénomination dans plusieurs textes et des tentatives de définition ont été faites au cours des années (*ibidem*).

En revanche, comme le remarque G. T., il est difficile de trouver une référence au terme « minorité » dans le contexte juridique français. En effet, en droit interne ce concept se heurte aux valeurs constitutionnelles de la République. L'exemple que cite l'étudiante à ce propos concerne tout particulièrement les minorités linguistiques et la question de la Charte européenne des langues régionales ou minoritaires qui a été signée en 1999 par la France, mais pas ratifiée. Le Conseil d'État avait alors annoncé « qu'en adhérant à la Charte la France méconnaîtrait les principes constitutionnels d'indivisibilité de la République, d'égalité devant la loi, d'unicité du peuple français et d'usage officiel de la langue française »[16]. Dans cet avis de refus, remarque l'étudiante, « le Conseil d'État choisit de parler de 'groupes de locuteurs de langues régionales ou minoritaires' et n'utilise

[16] L'avis du Conseil d'État du 30 juillet 2015 est disponible au lien suivant : https://mjp.univ-perp.fr/france/CE2015.htm (consulté le 18/07/2021).

même pas la dénomination de 'minorité linguistique' ». Cette décision a ensuite été confirmée par le Conseil d'État en 2013. À partir de cet exemple, G. T. a recherché dans la traduction automatique les équivalents en français de « *linguistic minority* » dans plusieurs dispositifs de traduction automatique. Ce qui en ressort, c'est tout d'abord le manque de variations : DeepL et Google Translate proposent seulement l'équivalent « minorité linguistique ». Des résultats similaires sont affichés lorsqu'on cherche le terme en contexte dans Linguee ou Reverso Context[17]. Il est intéressant de noter que tous les exemples proposés par le concordancier Linguee[18] sont tirés de sources canadiennes, européennes et internationales. Voyons-en quelques-uns dans le tableau 2 :

[...] availability of bilingual weather forecasts, the service is not always available in the language of the linguistic minority. Source : clo.gc.ca	[...] bilingue des prévisions météorologiques, le service n'est pas toujours disponible dans la langue de la minorité linguistique. Source : clo.gc.ca
They may be disabled or belong to an indigenous group or a linguistic minority. Source : unesdoc.unesco.org	Ils peuvent être handicapés ou appartenir à un groupe autochtone ou à une minorité linguistique. Source : unesdoc.unesco.org
Finland is a bilingual country in which the linguistic minority enjoys a high level of protection, as I should like to demonstrate. Source : europarl.europa.eu	La Finlande est un pays bilingue où la minorité linguistique fait l'objet d'une vraie protection, comme j'aimerais en faire la démonstration Source : europarl.europa.eu

Tableau 2 : Exemples des équivalents de « linguistic minority » de l'anglais au français dans Linguee

Le fait de ne pas trouver des contextes liés à des sources institutionnelles françaises s'explique sans doute par les raisons que nous venons d'exposer. Il est alors logique de ne pas repérer le terme « minorité linguistique » dans le discours institutionnel français, car il ne décrit aucune réalité qui puisse être conceptualisée dans le respect des principes républicains. En effet, la France considère, d'après l'article 2 de sa Constitution, le français comme (seule) langue de la République et utilise exclusivement dans ses textes officiels le terme « langue régionale » pour se référer à l'alsacien, au basque, breton, catalan, etc[19]. Toutefois, nous considérons que

[17] Tous ces dispositifs ont été consultés pour la dernière fois le 18/07/2021.

[18] Nous rappelons que Linguee permet de remonter à la source de l'extrait proposé.

[19] À ce propos, il est significatif de noter qu'en mai 2021 le Conseil Constitutionnel a censuré l'article 4 de la loi sur la protection patrimoniale des langues régionales et leur promotion en particulier concernant l'enseignement immersif. Cet article a été considéré comme contraire à l'article 2 de la Constitution que nous venons de mentionner. Plus d'informations à ce sujet sur le site de la République française au lien suivant : https://www.vie-publique.fr/en-bref/280872-langues-regionales-lenseignement-immersif-en-question (consulté le 18/07/2021).

l'absence de la variante diatopique du contexte français devrait être signalée dans les dispositifs de traduction d'autant plus que la non-reconnaissance du terme est un sujet qui soulève des débats, notamment dans les domaines politique et académique (Bassac *et al.* 2018).

Après avoir présenté les deux recherches menées par nos étudiant·e·s et qui sont représentatives du travail qu'il·elle·s ont fait en classe, nous allons maintenant nous pencher sur l'enquête empirique que nous avons effectuée en parallèle par la présentation de deux questionnaires en début et en fin de cours. Ce type de travail nous a permis surtout d'observer une prise de conscience critique de l'étudiant·e par rapport à l'usage des dispositifs de traduction automatique.

4. L'enquête statistique et l'évolution du premier au deuxième questionnaire

Les deux questionnaires, qui ont été présentés à tou·te·s les étudiant·e·s des Universités participant au Projet européen AI4EI, ont été élaborés par le groupe de recherche à partir du modèle qui a été présenté dans l'introduction de la deuxième partie de cet ouvrage. C'est la comparaison des données recueillies par les deux questionnaires du début et de la fin du cours qui nous a permis d'évaluer les compétences acquises par les participant·e·s pendant le travail didactique.

Parmi les résultats obtenus, plusieurs points méritent d'être explicités. Tout d'abord, nous avons remarqué à la fin du cours une capacité majeure à identifier l'erreur ou à évaluer les résultats de sortie. Par exemple, comme nous avons pu également le constater précédemment pour le terme « minorité » ou « *whistleblower* », les étudiant·e·s tendent à vérifier, quand cela est possible, la source de l'extrait proposé par les dispositifs de traduction et à analyser le contexte de parution du terme. C'est sans doute pour cette raison que, lors du deuxième questionnaire et contrairement à ce qui se passe au début du cours, il·elle·s considèrent que les recherches lancées dans les concordanciers peuvent être plus performantes par rapport aux résultats des traducteurs automatiques qui n'affichent pas le contexte. En effet, dans le premier questionnaire qu'on leur a demandé de remplir, 34 participant·e·s sur 38 considèrent les traducteurs automatiques très fiables ou assez fiables contre 28 qui déclarent que les plateformes bilingues (les concordanciers) sont très fiables ou assez fiables.

Cette tendance s'inverse au second semestre avec les chiffres suivants : 29 pour les traducteurs et 32 pour les concordanciers.

Deuxièmement, les étudiant·e·s constatent un aplatissement de la diversité linguistique et de la variation qui peut affecter surtout les usagers ne maîtrisant pas les langues concernées par la traduction ou les non-experts, dans le cas de la traduction spécialisée[20]. Il convient de noter que, si dans le premier questionnaire la réduction de la barrière linguistique est présentée comme le premier aspect positif qui permettrait de favoriser le dialogue entre civilisations (16 participant·e·s sur 38 mettent en avant cet aspect), dans le questionnaire de fin de cours cet argument devient secondaire. Au contraire, dans le deuxième questionnaire, il·elle·s soulignent les imprécisions fréquentes lorsque la traduction concerne des langues non romanes et non germaniques — surtout le russe, le chinois, le coréen — et, par conséquent, un manque de représentativité de certaines langues-cultures.

En ce qui concerne plus particulièrement la variation terminologique, la totalité des étudiant·e·s déclarent au deuxième questionnaire que l'inexactitude du résultat risque d'entraîner des conséquences négatives sur la compréhension d'un concept juridique, politique et/ ou social[21].

Troisièmement, les participant·e·s évoquent toujours la question de la délégation de la mémoire vers les dispositifs de traduction[22]. En effet, celle-ci est présentée à la fois comme un aspect positif (presque exclusivement dans le premier questionnaire) et négatif. L'outil est parfois perçu comme un objet qui permet d'obtenir des résultats rapidement et sans effort, alors que, dans la plupart des cas, plusieurs participant·e·s soulignent le risque de dépendance et d'appauvrissement qui entraînerait à long terme l'usage des dispositifs en tant que « facilitateurs » d'apprentissage des langues étrangères.

Enfin, nous avons pu constater l'évolution de la perception du rôle de l'humain par rapport à la traduction automatique. Si dans le premier ques-

[20] Ces données relèvent du troisième bloc (*attitudinal questions*), qui prévoyait des questions à réponses ouvertes.

[21] Nous précisons que la question a été formulée comme suit : « La traduction erronée d'un terme pourrait-elle entraîner des conséquences négatives sur la compréhension d'un concept juridique, politique et/ ou social ? ». Elle a été insérée seulement dans le deuxième questionnaire.

[22] Sur la délégation de la mémoire aux outils numériques voir Emmanuël Souchier (2004), qui traite la question de la progressive externalisation de la mémoire humaine, individuelle et collective vers l'outil technologique.

tionnaire, parmi les questions du troisième bloc à réponses ouvertes, sept personnes pointent le risque de voir disparaître la profession des linguistes et des traducteurs humains, ces inquiétudes disparaissent dans le deuxième questionnaire et sont remplacées par l'intérêt pour l'interaction personne-machine : les étudiant·e·s évoquent la facilité pour l'humain de comprendre rapidement un contexte, combinée à la nécessité de procéder à des recherches approfondies et à la post-édition lors de la traduction automatique.

En définitive, l'enquête par questionnaire que nous avons menée nous a permis d'une part, d'évaluer les connaissances d'un public universitaire par rapport à l'usage des dispositifs de traduction automatique et, d'autre part, de dresser un bilan de notre proposition pédagogique interdisciplinaire, comme nous allons le préciser dans les conclusions.

Conclusion

Les analyses menées en classe par nos étudiant·e·s, ainsi que les données recueillies par les questionnaires soumis lors de notre enquête pédagogique, montrent que les étudiant·e·s s'intéressent aux dénominations plurielles du même concept, telles qu'elles ressortent des différents contextes d'usage, notamment politique et juridique. Cet intérêt tend à développer l'esprit critique des étudiant·e·s par rapport aux résultats de sortie proposés par les dispositifs de TA.

À propos de notre questionnement initial, à savoir quelle est l'interaction entre la terminologie multilingue, le droit et les dispositifs de traduction automatique, il convient de souligner que les algorithmes intelligents risquent de réduire les variations linguistiques et de susciter des interprétations discriminatoires (Bartoletti 2020). En ce sens, ces dispositifs ne peuvent pas être considérés comme de simples outils, mais comme des « prothèses » du langage et de la communication (Perea, Wagener 2020 : 6), qui contribuent à façonner et à perpétuer la réalité qu'ils construisent, d'autant plus que les évolutions du numérique dans le droit amènent à s'interroger actuellement sur plusieurs défis, comme la portée des évaluations prédictives et les effets de la machine sur l'interprétation juridique.

Ce type d'expérimentation didactique, au carrefour des langues et du droit, nous semble donc apporter des éléments de réflexion qui peuvent contribuer à mieux comprendre les conséquences des technologies de l'intelligence artificielle sur la variation terminologique et sur l'expertise juridique.

Bibliographie

Abou Ramadan Moussa (2016). « La définition des minorités en droit international ». Dans : Lionel Obadia, Anne-Laure Zwilling (éds). *Minorité et communauté en religion*. Strasbourg : Presses universitaires de Strasbourg, 79-98.

Bartoletti Ivana (2020). *An Artificial Revolution: On Power, Politics and AI*. Londres : The Indigo Press.

Bassac Christian, Busquets Joan, Guset Victor, Pascaud Antoine, Viaut Alain (2018). « Pour une définition de la notion de minorité linguistique : les difficultés du vague ». *Lengas*, 83, en ligne. URL : http://journals.openedition.org/lengas/1713.

Candel Danielle (2004). « Wüster par lui-même ». Dans : Colette Cortès (éd). *Des fondements théoriques de la terminologie*. Paris : Université Paris Diderot : Cahiers du CIEL, 15-31.

Dufour Françoise, Rosier Laurence (2012). « Introduction. Héritages et reconfigurations conceptuelles de l'analyse du discours 'à la française' : perte ou profit ? ». *Langage et Société*, 140, 5-13.

Gréciano Philippe, Humbley John (éds) (2011). « Langue et droit : terminologie et traduction ». *Revue française de linguistique appliquée*, XVI/1. Paris : Publications linguistiques.

Guidère Mathieu (2008). *La communication multilingue. Traduction commerciale et institutionnelle*. Bruxelles : De Boeck.

Humbley John (2018a). « La terminographie entre langue et discours : réflexions historiques et épistémologiques ». Dans : Jana Altmanova, Maria Centrella, Katherine E. Russo (éds). *Terminology & Discourse / Terminologie et discours*. Berne : Peter Lang, 69-92.

Humbley John (2018b). « L'onomasiologie comme principe constituant de la néonymie diachronique ». *ELAD-SILDA*, 1, en ligne. URL : https://publications-prairial.fr/elad-silda/index.php?id=272.

L'Homme Marie-Claude (2004). *La terminologie : principes et techniques*. Montréal : Presses de l'Université de Montréal.

Lagarde Marcela (2006). « Del femicidio al feminicidio ». *Desde el Jardín de Freud*, 6, en ligne, 216-225. URL : https://revistas.unal.edu.co/index.php/jardin/article/view/8343/8987.

Lagarde Marcela (2008). « Violencia feminicida y derechos humanos de las mujeres ». In : Margaret Bullen, Carmen Diez Mintegui (éds). *Retos teóricos y nuevas prácticas*. Mexique : Universidad Autónoma de México, 209-239.

Lapalus Marylène (2015). « Feminicidio / femicidio : les enjeux théoriques et politiques d'un discours définitoire de la violence contre les femmes ». *Enfances, Familles, Générations*, 22, 85–113.

Maingueneau Dominique (2002). « Positionnement ». Dans : Patrick Charaudeau, Dominique Maingueneau (éds). *Dictionnaire d'analyse du discours*. Paris : Seuil, 453-454.

Pascual Edmond (2004). *La communication écrite en diplomatie*. Perpignan : Presses universitaires de Perpignan.

Perea François, Wagener Albin (2020). « 'Envisagez de retirer ce terme sensible'. Correction et prescription automatique du langage non discriminant », *Communication & langages*, 206/4, 3-21.

Pulcinelli-Orlandi Eni (1996). *Les formes du silence. Dans le mouvement du sens*. Paris : Éditions de cendres.

Radford Jill, Russell Diana E.H. (1992). *Femicide: The Politics of Woman Killing*. New York : Macmillan.

Raus Rachele (2013). *La terminologie multilingue. La traduction des termes de l'égalité H/F dans le discours international*. Bruxelles : De Boeck.

Sager Juan C. (1990). *A Practical Course in Terminology Processing.* Amsterdam/ Philadelphia : John Benjamins.

Souchier Emmanuël (2004). « Mémoires - outils - langages. Vers une « société du texte » ? ». *Communication & langages*, 139. Dossier « Le « constructivisme », une nouvelle vulgate pour la communication ? », 41-52.

Zanola Maria Teresa (2018). *Che cos'è la terminologia*. Rome : Carocci.

Zanotta Machado Lia (2019). « Féminicide : nommer pour exister », *Brésil(s)*, 16, en ligne. URL : http://journals.openedition.org/bresils/5576