hybrid
Revue des arts et médiations humaines
Journal of arts and Human Mediations

Revue *Hybrid*, n° 7

Le réseau créatif des langu.ages

Intelligence artificielle ET/OU diversité linguistique : les paradoxes du traitement automatique des langues

Claire Larsonneur

Claire Larsonneur est maître de conférences hors classe à l'Université Paris 8 au sein de l'équipe TransCrit. Ses travaux portent sur la littérature numérique, le contexte économique, technologique et social de la traduction et les humanités numériques. Elle a codirigé le projet Labex Arts-H2H « Le sujet digital », le colloque de Cerisy Posthumains et subjectivités numériques et le numéro d'*Angles* « Digital Subjectivities ». Elle fait partie de l'équipe du projet Épistémologies et pratiques des humanités numériques (Paris 8, UVSQ, projet OPERAS). Son dernier ouvrage en date, *When Translation Goes Digital*, codirigé avec Renée Desjardins (Université de Saint-Boniface, Canada) et Philippe Lacour (Université de Brasilia, Brésil) est paru fin 2020 chez Palgrave Macmillan.

Résumé

Les technologies de traitement automatique des langues remodèlent le paysage linguistique d'Internet de manière variée et complexe. Les outils gratuits de traduction neuronale, comme Google Translate et DeepL, tout comme les agents intelligents tels que Siri et Alexa, facilitent la production et la diffusion de contenus dans de nombreuses langues, favorisant ainsi le multilinguisme des interfaces. Il y a plus de langues en ligne, elles sont plus visibles et les contenus sont plus accessibles. Dans le détail toutefois, on observe une érosion de la diversité

linguistique : en effet ces outils tendent à standardiser et à homogénéiser les expressions, ils sont basés sur un nombre restreint de corpus et ils valorisent une performance en matière de taux d'engagement des utilisateurs plutôt que de qualité des contenus.

Mots clés

traitement des langues naturelles, diversité, multilinguisme, intelligence artificielle, IA, agents intelligents

Traduction de l'anglais par Armelle Chrétien

Mise en ligne : 15 juin 2021

D'après les projections établies par une équipe de chercheurs en intelligence artificielle du Future of Humanity Institute (Institut pour l'avenir de l'humanité), d'ici 2024 les programmes informatiques pourront produire des traductions équivalentes en qualité aux traductions humaines et d'ici 2049 ils pourront écrire un *bestseller* digne de figurer sur la liste du *New York Times*[1]. Ces projections se fondent sur les remarquables avancées récemment réalisées en traitement automatique des langues (TAL) grâce au perfectionnement des technologies d'intelligence artificielle (IA), à une puissance de calcul accrue et à d'immenses bases de données collectées principalement sur le Web. Si l'attribution du Booker Prize à un robot tient encore de la science-fiction, nombre de textes du quotidien – des modes d'emploi aux bulletins météorologiques – relèvent désormais d'une production industrielle. La traduction automatique est aujourd'hui le meilleur exemple de technologies de traitement automatique des langues devenues monnaie courante sur de grandes plateformes comme celles d'Amazon, Twitter, Facebook et Google.

Le nombre de langues disponibles en ligne semble actuellement en forte hausse. On attend désormais des grandes entreprises et des institutions publiques qu'elles soient accessibles en ligne dans plusieurs langues et le multilinguisme n'est plus l'exception. Par exemple, le site Web des Témoins de Jéhovah est actuellement accessible en 750 langues, ce qui lui vaut le titre de « site le plus traduit au monde ». Wikipédia est disponible en 287 langues, Google en 149[2]. Facebook peut être consulté en plus d'une centaine de langues et se targue d'effectuer plus de 6 milliards de traductions par jour à partir de 4 000 paires de langues[3]. La

---

1 Charlotte Edmond, « This is when a robot is going to take your job, according to Oxford University », 26 juillet 2017. [En ligne] https://www.weforum.org/agenda/2017/07/how-long-before-a-robot-takes-your-job-here-s-when-ai-experts-think-it-will-happen/ [consulté le 18 septembre 2019].

2 Mishel Shaji, « What is the most translated website in the world ? », 17 mars 2019. [En ligne] https://www.quora.com/What-are-the-top-three-most-translated-Websites-and-how-many-languages-do-they-each-accommodate [consulté le 18 septembre 2019].

3 Facebook does not specify whether the impressive figure of 6 billion translations refers to the number of char acters, the number of words or the number of documents : Khari Johnson, « Facebook Messenger launches translations by intelligent assistant M » 1er mai 2018. [En ligne] https://venturebeat.com/2018/05/01/facebook-messenger-launches-translations-by-intelligent-assistant-m/ [consulté le 18 septembre 2019].

mondialisation a accru les besoins en traduction et en localisation à travers le monde : le poids du marché mondial des services en langue a ainsi doublé au cours des 16 dernières années pour atteindre 46,9 milliards de dollars en 2019[4], et le marché mondial de la traduction automatique devrait connaître une croissance de 19 % entre 2020 et 2024[5]. Le boom de la demande en traduction alimente la demande en outils de TAL capables de fournir des solutions instantanées et à bas coût. En retour, une plus grande efficacité dans le traitement automatique des langues permet de créer et de diffuser plus de contenus dans plus de langues. Une boucle de rétroaction opèrerait ainsi entre ces technologies et le multilinguisme en ligne.

Mais l'augmentation de contenus multilingues, sur le Web et en dehors, ne signifie pas qu'il y ait plus de personnes multilingues, définies comme maniant trois langues ou plus au quotidien[6]. Au contraire, celui qui accède plus facilement à des contenus dans sa langue maternelle peut être moins motivé pour apprendre d'autres langues. Le multilinguisme coïncide-t-il avec la diversité linguistique ? Certes les plateformes en ligne se vantent de proposer des contenus en de nombreuses langues et cela fait partie de leurs stratégies de communication. Mais l'Union européenne a récemment souligné le risque d'extinction numérique qui guette plus de 20 langues européennes, moins parlées et insuffisamment représentées sur le Web, dans sa résolution du 11 septembre 2018 sur « L'égalité des langues à l'ère numérique ». Selon John Paolillo, Internet favorise clairement les langues dominantes et les normes techniques qui leur sont associées. Son impact sur les langues minoritaires ou minorisées est moins évident : il peut les affaiblir comme les renforcer[7]. Enfin, il est permis de s'interroger sur les représentations et les valeurs associées au TAL. Manohar Paluri, responsable de l'IA chez Facebook, affirmait ainsi en mai 2019 que l'intelligence artificielle « nous fournit un outil puissant pour résoudre les problèmes de langue sans être versé en langues », tandis que Google Translate annonçait en juillet 2019 se tourner vers un traitement automatique neuronal massivement multilingue. Comment la même technologie peut-elle être massivement multilingue et faire fi de la linguistique ?

---

4 Elena Mazareanu, « Market size of the global language services market » 9 août 2019. [En ligne] https://www.statista.com/statistics/257656/size-of-the-global-language-services-market/ [consulté le 3 juillet 2020].

5 Technavio, « Globalization of business to boost growth », [En ligne] https://www.businesswire.com/news/home/20200529005034/en/Machine-Translation-Market-2020-2024-Globalization-Business-Boost 29 mai 2020 [consulté le 3 juillet 2020].

6 Charlotte Kemp, « Defining multilingualism », in Larissa Aronin and Britta Hufeigein (dir.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009, p. 14.

7 John Paolillo, « Language diversity on the Internet : examining linguistic bias », in *Measuring Linguistic Diversity on the Internet* (UNESCO Institute for Statistics), Paris, UNESCO, 2005, p. 55.

Dans l'esprit de ce qu'Yves Citton appelle l'« archéologie des média[8] », c'est à partir de l'étude des caractéristiques techniques du TAL que je vais aborder la question de son influence sur le multilinguisme et la diversité linguistique. Sommes-nous en train d'assister à l'émergence d'une économie mondiale plus multilingue mais moins linguistiquement diverse ? Dans le domaine du traitement automatique des langues, l'intelligence artificielle signerait-elle la fin de la diversité ? Après avoir précisé la distinction entre multilinguisme et diversité linguistique, je me pencherai sur le cas spécifique de la traduction automatique neuronale avant d'étudier l'impact du TAL en général sur la diversité linguistique à plusieurs niveaux.

Multilinguisme *vs* diversité linguistique

« Les 10 millions de sites Web les plus fréquentés sont rédigés dans un panel de moins de 10 langues pour 90 % d'entre eux : seule une infime fraction des 7 097 langues répertoriées domine le Net[9] », constate Emma Charlton du Forum économique mondial. Si la part historiquement prépondérante de l'anglais sur le Web diminue aujourd'hui en raison d'une croissance rapide des contenus publiés en arabe, russe ou chinois, établir une cartographie des langues en ligne demeure une tâche ardue. Internet World Stats étudie ainsi le nombre d'utilisateurs d'Internet par langue et le taux de pénétration d'Internet au sein d'une population donnée[10]. Une telle approche montre d'importants écarts entre le nombre de locuteurs natifs et le nombre d'utilisateurs d'Internet : seuls 53 % des arabophones utilisent Internet, ou 35,2 % des francophones, par comparaison avec l'allemand ou le japonais où le taux de pénétration est de 93,8 %. Il arrive aussi qu'une même personne se serve de plusieurs langues sur le Web (par exemple l'anglais et le coréen), à différents degrés et selon les contextes.

---

8 Yves Citton, *Médiarchies*, Paris, Seuil, 2017, p. 194 : « L'archéologie des media, c'est surtout une autre façon d'approcher dans le long terme la matérialité physique des modes de communication qui conditionnent tout à la fois nos organisations collectives et nos représentations mentales. C'est dans le fonctionnement matériel, généralement caché, des appareils qu'il convient d'aller chercher la raison des qualités occultes qui en émanent. L'hypothèse de départ est que nos appareils en savent davantage que nous-mêmes sur ce qu'ils font de nous quand nous croyons nous servir d'eux. »

9 Emma Charlton, « The Internet has a language diversity problem », World Economic Forum, 13 décembre 2018. [En ligne] https://www.weforum.org/agenda/2018/12/chart-of-the-day-the-internet-has-a-language-diversity-problem/ [consulté le 18 septembre 2019].

10 Internet World Stats. [En ligne] https://www.internetworldstats.com/stats7.htm [consulté le 22 juin 2020].

Une telle complexité linguistique fait écho à la multiplicité et la connexité des définitions du multilinguisme, comme le rappelle Charlotte Kemp. Le terme « multilingue » renvoie en général à des individus parlant trois langues ou plus. Kemp distingue cette notion de celle de polyglossie, « un terme généralement employé en sociolinguistique pour désigner des communautés dans lesquelles différentes langues ou variétés de langues sont employées par une partie ou l'ensemble des membres d'une communauté particulière dans laquelle ils assument différents rôles[11] ». Cependant, il faut souligner que ni l'une ni l'autre de ces définitions ne prennent en compte les spécificités de la production et de la circulation linguistique en ligne : certaines personnes, alors même que leur vie hors ligne se déroule dans un environnement essentiellement monolingue (par exemple en France), peuvent ainsi appartenir à une communauté en ligne bilingue, comme celles du *gaming* ou des logiciels *open source*, où l'usage de l'anglais est très répandu. Envisager les communautés en ligne comme des « espaces d'affinité translocaux[12] » permet de prendre ses distances avec la cartographie des langues hors ligne, souvent alignée sur des territoires physiques et politiques, ainsi qu'avec une évaluation strictement quantitative. S'intéresser à l'inverse aux pratiques linguistiques et à leur complexité peut s'avérer un axe de recherche plus fécond.

Dans le cadre de cette recherche, je propose d'établir une distinction entre le multilinguisme en ligne, défini par des indicateurs de quantité, de visibilité et d'accessibilité (tels que le nombre de sites Web disponibles dans une langue donnée, ou le nombre de langues proposées par un site), et la diversité linguistique, caractérisée par des pratiques et des usages, des niveaux de compétence, des degrés d'intelligibilité mutuelle et des questions d'identité[13]. Cette distinction suivra l'axe proposé par Peter Strevens[14] :

> L'un des principaux problèmes rencontrés par l'étude des langues est de réconcilier une fiction nécessaire et opportune avec un vaste ensemble de faits inopportuns. Cette fiction est celle de « la langue » – anglaise, chinoise, navajo, kashmiri. Les faits résident quant à eux dans l'immense diversité manifestée dans l'usage réel qu'en font les individus lorsqu'ils emploient une langue donnée.

---

11 Charlotte Kemp, « Defining multilingualism », in Larissa Aronin and Britta Hufeigein (dir.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009, p. 15.

12 Sirpa Leppänen and Saija Peuronen, « Multilingualism on the Internet », in Marilyn Martin-Jones, Adrian Blackledge et Angela Creese (dir.), *The Routledge Handbook of Multilingualism*, London/New York, Routledge, 2015, p. 390.

13 Charlotte Kemp, « Defining multilingualism », in Larissa Aronin et Britta Hufeigein (dir.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009, p. 23.

14 Peter Strevens cité par Charlotte Kemp, « Defining multilingualism », in Larissa Aronin et Britta Hufeigein (dir.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009, p. 16.

Dans le cadre d'une analyse des politiques linguistiques en ligne mises en place par les grandes entreprises, il peut être utile d'établir une distinction entre multilinguisme et diversité linguistique. Ainsi, le multilinguisme en ligne des entreprises ne reflète bien souvent pas la réalité linguistique hors ligne, soit en raison de contraintes techniques, soit de choix marketing. En septembre 2019, la société Zara est ainsi physiquement implantée dans 96 pays allant de l'Azerbaïdjan à l'Équateur ; sa présence en ligne recouvre 202 marchés[15]. Sa stratégie de localisation est clairement définie : depuis 2016, elle couvre toutes les langues de l'UE et son site suisse est disponible en 4 langues[16]. Cependant, le principal site de la société-mère Inditex n'est disponible qu'en anglais et en espagnol. De plus, les clients sont automatiquement redirigés vers le site Web du pays où sont répertoriées les adresses IP de leurs ordinateurs sans pouvoir accéder au contenu propre à d'autres marchés ou d'autres langues. La présence linguistique de Zara en ligne est incontestablement multilingue mais on ne peut pas dire qu'elle promeuve la diversité des langues du point de vue de l'internaute ou du client.

Outre les sites Web, on peut également étudier les politiques linguistiques à l'œuvre dans la conception des assistants personnels intelligents. L'hétérogénéité de leurs paramètres linguistiques est révélatrice de la diversité des enjeux et des stratégies adoptées par les entreprises[17]. Au mois de juin 2020, Cortana était disponible en 7 langues. L'assistant n'opère aucune distinction entre les variantes australienne, britannique ou américaine de l'anglais, ni entre les variantes mexicaine et espagnole, la seule région bilingue étant le Canada (anglais/français). Chez Apple, Siri est davantage multilingue. En septembre 2019, les paramètres de Siri sur un ordinateur Mac proposent un choix de 41 langues dont certaines sont localisées : 9 variantes de l'anglais, dont une pour Singapour, 4 variantes du chinois, mais une seule en arabe, alors que les normes ISO dénombrent plus de 30 variétés de l'arabe (section 639-3). On retrouve une grande hétérogénéité dans les paramètres de genre : Siri sera incarné au choix par une voix de femme ou d'homme dans les variantes américaine, britannique et australienne de l'anglais, mais les variantes irlandaise et sud-africaine sont

15 [En ligne] https://www.inditex.com/about-us/our-brands/zara [consulté le 22 juin 2020].

16 *Internet retailing*, « Zara. Communication without borders », 7 août 2017. [En ligne] https://internetretailing.net/research-articles/research-articles/zara-communication-without-borders [consulté le 17 décembre 2019].

17 Globalme.com, « Language support in voice assistants compared » (mis à jour 2019), 28 novembre 2019. [En ligne] https://www.globalme.net/blog/language-support-voice-assistants-compared [consulté le 17 décembre 2019].

uniquement disponibles dans la version féminine. L'assistant Alexa d'Amazon est proposé en sept langues mais uniquement au travers de voix féminines[18].

De telles disparités dans les politiques de langues des entreprises résultent d'un enchevêtrement complexe de logiques distinctes : l'identification de marchés prometteurs, la disponibilité des données, le degré de connaissances linguistiques des décideurs, les coûts entraînés par l'ajout d'une nouvelle langue, etc. Les paramètres linguistiques des plus gros éditeurs de logiciels, majoritairement basés aux États-Unis, sont logiquement conçus et mis en application en anglais américain puis, dans un second temps, localisés dans d'autres langues correspondant à des marchés porteurs. Pour Cortana (Microsoft), les langues sont distribuées en 13 « régions », une catégorie maison qui n'est calquée ni sur les pays ni sur les États, et pas même sur les marchés. Ses services ne permettent pas de basculer d'une langue à l'autre : « Si vous modifiez votre région, vous ne pourrez peut-être pas effectuer d'achats dans le Microsoft Store, ni utiliser certains éléments que vous avez achetés, comme des adhésions ou des abonnements, des jeux, des films ou des programmes TV[19]. » En dépit du multilinguisme de Cortana, on observe une fois de plus que l'internaute n'a pas accès à une grande diversité linguistique. Outre les critères économiques décrits ci-dessus (coûts, données, etc.), les politiques linguistiques des entreprises dépendent de représentations culturelles. Luke Munn a ainsi montré en quoi la restriction initiale des paramètres de voix d'Alexa à une seule version féminine était ancrée dans des stéréotypes historiques de la standardiste, une servante à la voix douce[20]. Siri, à l'inverse, est présenté par Apple comme un concierge d'hôtel ou un majordome : l'importance de l'aisance conversationnelle dans les interactions peut expliquer un choix plus étendu de variantes linguistiques et d'options de voix.

L'hétérogénéité du paysage linguistique des sites et des interfaces des plateformes en ligne semble ainsi favoriser les langues dominantes. Pour explorer plus avant le multilinguisme et la diversité linguistique en TAL, il nous faut étudier plus en détail les technologies d'IA utilisées, à commencer par le cas de la traduction automatique neuronale.

---

18 Un plus large choix de variantes linguistiques et de voix a été introduit en octobre 2018 lorsqu'Alexa a intégré une application de synthèse vocale, Amazon Polly : B. J. Haberkorn, « Amazon Polly voices in Alexa skills now generally available », 23 octobre 2018. [En ligne] https://developer.amazon.com/fr/blogs/alexa/post/baee53c1-5b03-4580-b57a-ee9510413354/amazon-polly-voices-in-alexa-skills-now-generally-available [consulté le 17 décembre 2019].

19 [En ligne] https://support.microsoft.com/en-ie/help/4026948/cortanas-regions-and-languages (mis à jour pour la dernière fois le 27 mai 2020) [consulté le 22 juin 2020].

20 Luke Munn, « Alexa and the intersectional interface », *Angles. New Perspectives on the Anglophone World*, n° 7 « Digital Subjectivities », juin 2018. [En ligne] https://angles.edel.univ-poitiers.fr:443/angles/index.php?id=1492 [consulté le 22 juin 2020].

Le cas de la traduction automatique neuronale

La traduction automatique neuronale (TAN), développée grâce à l'intelligence artificielle, offre aujourd'hui des résultats fluides, convaincants et dont la qualité ne cesse de s'améliorer. Certains outils grand public de traduction neuronale offerts en accès libre, tels que Google Translate ou DeepL, associent des technologies de traduction automatique statistique et d'apprentissage profond. La méthode statistique se sert d'immenses corpus de segments traduits pour faire correspondre un texte source au texte cible le plus probable : comme la traduction retenue est la combinaison la plus probable, le procédé favorise les occurrences les plus courantes au détriment de la rareté. Les algorithmes d'IA sont eux aussi entraînés à partir de vastes corpus bilingues voire multilingues ce qui leur permet d'« apprendre » à prédire la suite la plus probable pour un début de phrase donné. Les résultats actuels en TAN sont suffisamment concluants pour venir concurrencer la traduction humaine : dans des domaines spécialisés, les textes cibles produits par traduction automatique neuronale sont aujourd'hui généralement considérés « adéquats[21] ». Les technologies d'IA à l'œuvre dans la traduction automatique neuronale sont très proches de celles utilisées dans d'autres champs où le savoir-faire et l'agir humains ont longtemps été considérés comme essentiels : diagnostiques médicaux, surveillance des cultures agricole, voitures ou drones autonomes, etc.[22] En effet toutes ces technologies recourent à de vastes bases de données, à la reconnaissance de formes et à l'apprentissage profond. Elles sont mises au point par les mêmes équipes, comme l'illustre la trajectoire professionnelle de M. Olszewski : l'ancien responsable d'Amazon Translate a rejoint Uber en 2019 pour développer des véhicules autonomes à partir de technologies proches[23]. Enfin les algorithmes décisionnels ont encore un point commun : ils ont été conçus pour être intégrés à de multiples applications informatiques et même à des appareils physiques tels que des voitures, des réfrigérateurs, des vêtements « intelligents » ou des assistants domestiques, comme Echo (Amazon) ou Home Hub (Google).

Après ce rapide tour d'horizon des technologies d'IA, on peut se recentrer sur la traduction neuronale et son influence sur le paysage linguistique. Bien que l'objectif de la TAN soit de permettre des traductions rapides à partir de nombreuses paires de langues, accroissant

---

21 Joss Moorkens, Sheila Castilho, Federico Gaspari et Stephen Doherty (dir.), *Translation Quality Assessment. From Principles to Practice*, Cham, Springer International Publishing, 2018.

22 Cedric Villani, *For a Meaningful Artificial Intelligence. Towards a French and European Strategy*, mars 2018. [En ligne] https://www.aiforhumanity.fr/en/ [consulté le 20 septembre 2019].

23 Marion Marking, « Alon Lavie joins Unbabel. He's not the only exec who recently left Amazon's machine translation group », *Slator*, 18 avril 2019. [En ligne] https://slator.com/people-moves/alon-lavie-joins-unbabel-hes- not-the-only-exec-who-recently-left-amazons-translation-group/ [consulté le 22 juin 2020].

de facto le multilinguisme à l'échelle globale, je suggère qu'elle le fait aux dépens de la diversité linguistique. Premièrement et à l'instar des autres technologies d'IA, la traduction automatique neuronale entraîne une perte d'information. C'est que démontre Pasquinelli, sur la base de trois arguments : les limites des corpus d'entraînement (car il s'agit toujours d'échantillons), l'utilisation de récurrences, de catégories et de taxinomies, et l'existence de biais. La déperdition est particulièrement notable pour les métaphores, les mots rares et les néologismes, qui sont des anomalies statistiques.

Pasquinelli souligne que l'IA est avant tout une technique de compression d'information et non la manifestation d'une cognition non humaine. Il montre aussi combien la conception de ces outils et de leurs usages est contrainte par des impératifs économiques :

> Les algorithmes ont toujours été des dispositifs de nature économique destinés à livrer un résultat à partir du moins d'étapes et de la plus faible consommation de ressources possible : espace, temps, énergie, etc. La présente course à l'armement entre sociétés d'IA se résume toujours à la découverte de l'algorithme le plus rapide pour calculer les modèles statistiques. La compression de l'information mesure par conséquent le taux de profit de ces entreprises, mais aussi le taux de perte d'information. Or, cette perte se traduit souvent par une perte au niveau de la diversité culturelle à l'échelle mondiale[24].

Deuxièmement, la logique même de la traduction automatique neuronale revient à s'affranchir des contraintes des langues naturelles en les encodant sous forme de représentations mathématiques. À partir d'immenses corpus de segments bilingues et monolingues, les ingénieurs établissent une cartographie sémantique des relations entre les termes. Dans l'apprentissage bilingue supervisé, les expressions propres à la langue naturelle du texte source sont encodées sous forme de vecteurs sémantiques au sein de cette carte avant d'être décodées dans le texte cible. Les algorithmes sélectionnent les correspondances les plus probables sur la base de leur récurrence statistique au sein du corpus[25]. La TAN repose ainsi sur une double rupture avec la logique des langues naturelles : en introduisant un processus d'encodage et de décodage, et en opérant une sélection statistique.

Cette tendance a été encouragée par les derniers développements de la recherche en TAN, à savoir la traduction automatique massivement multilingue non supervisée. Au lieu de

---

24 Matteo Pasquinelli, « How a machine learns and fails — A grammar of error for artificial intelligence », *Spheres*, n° 5, « Spectres of AI », novembre 2019. [En ligne] http://spheres-journal.org/how-a-machine-learns-and-fails-a-grammar-of-error-for-artificial-intelligence/ [consulté le 13 décembre 2019].

25 Dorothy Kenny, « Machine translation », in Piers Rawling and Philip Wilson (dir.), *The Routledge Handbook of Translation and Philosophy*, Milton Park, Routledge, 2018, p. 428-445.

travailler à partir de paires de langues, les chercheurs de Google et de Facebook ont bâti des espaces de représentation communs à un grand nombre de langues. Facebook entraîne son modèle multilingue à partir de 93 langues, 30 familles de langues et 22 systèmes d'écriture[26]. Le modèle de Google prend en charge 103 langues entraînées à partir de 25 milliards d'exemples obtenus sur le Web en indexant des données et en extrayant des phrases parallèles. En 2019, les équipes d'IA de Facebook ont ainsi remporté un concours de traduction automatique anglais-birman en utilisant des techniques de rétrotraduction non supervisée[27]. Ces technologies impliquent toutefois un compromis entre transfert et interférence en vertu duquel la performance est améliorée pour des langues peu dotées et dégradée pour les langues bien dotées en raison de cette interférence et de capacités restreintes[28]. Mais les investissements massifs dans la recherche en TAN par Google, Facebook, Apple, Amazon et leurs équivalents asiatiques promettent des améliorations rapides et notables[29]. À ceci près que les outils ainsi développés ne visent pas seulement et peut-être même pas prioritairement à améliorer le traitement des textes proprement dits : l'objectif est que les machines soient capables d'articuler de manière fluide l'apprentissage de représentations, la compréhension de contenus, les systèmes de dialogue, l'extraction d'information, l'analyse d'opinion, la rédaction de résumés, la collecte et le nettoyage de données, et l'articulation voix/texte/image[30]. La TAN fait partie d'une palette de technologies de traitement des langues qui comprend l'aide à l'écriture (Microsoft Ideas), les outils

---

26 Gino Diño, « Facebook says it now has a powerful tool to think about language problems in a language agnostic way », *Slator*, 6 mai 2019. [En ligne] https://slator.com/technology/at-f8-facebook-says-it-now-has-a-powerful-tool-to-think-about-language-problems-in-a-language-agnostic-way [consulté le 18 septembre 2019].

27 Philippe Pajot, « La traduction automatique s'attaque aux langues rares », *La Recherche*, n° 554, décembre 2019, p. 20 *sq*.

28 Esther Bond, « What's so massive about Google's massively multilingual neural machine translation? », *Slator*, 18 juillet 2019. [En ligne] https://slator.com/technology/whats-so-massive-about-googles-massively-multilingual-neural-machine-translation/ [consulté le 18 septembre 2019].

29 Claire Larsonneur, « The disruptions of neural machine translation », *Spheres*, n° 5, novembre 2019. [En ligne] http://spheres-journal.org/the-disruptions-of-neural-machine-translation/ [consulté le 17 décembre 2019].

30 Facebook a créé un consortium qui finance de nombreux projets de recherche en TAL et Amazon a investi 200 millions de dollars sur des outils de reconnaissance vocale : Kyle Wiggers, « Facebook founds AI Language Research Consortium to solve challenges in natural language processing », 28 août 2019. [En ligne] https://venturebeat.com/2019/08/28/facebook-founds-ai-language-research-consortium-to-solve-challenges-in-natural-language-processing/ [consulté le 17 décembre 2019].

pédagogiques d'évaluation et d'annotation (Criterion), les assistants personnels intelligents (Siri, Alexa, Cortana), les agents conversationnels et bien d'autres. Bien que ces outils contribuent incontestablement à un développement du multilinguisme en ligne, ni la diversité linguistique ni même la qualité des échanges linguistiques (structurés, complexes, lexicalement riches) ne sont des enjeux prioritaires. Dans la section qui suit, nous étendrons le champ d'investigation au-delà de la traduction et prendrons en compte l'ensemble des technologies de la langue.

Limites de la diversité en TAL

Mieux comprendre comment la diversité linguistique est enrichie ou appauvrie par le traitement automatique des langues requiert de prendre en compte les différentes facettes de l'écosystème technique et économique des industries de la langue : les techniques employées, les indicateurs de performance, l'interaction entre l'humain et la machine, les dernières évolutions du marché et les questions de gouvernance.

La qualité du contenu généré par les outils de traitement de la langue dépend directement de la taille et de la teneur du corpus sur lequel les algorithmes sont entraînés. Dans le cas d'Alexa, l'objectif est d'alimenter des conversations plausibles et engageantes avec les utilisateurs : il faut pour cela mobiliser des modèles de détection de thèmes, des technologies de génération d'énoncés et d'annotation sémantique. Leurs développeurs ont donc travaillé à partir de corpus journalistiques, comme celui du *Washington Post*, de données extraites des réseaux sociaux comme Twitter ou Reddit et de bases de connaissance telles qu'Evi (Amazon), Freebase, Wikidata, Microsoft Concept Graph, Google Knowledge Graph et IMDB[31]. Or toutes ces sources d'information sont américaines, ce qui restreint le champ à une seule variante d'une langue ; en outre, elles sont loin de couvrir l'ensemble des types d'interactions linguistiques entre locuteurs réels. Elles ne reflètent donc pas toute la diversité des échanges anglophones à travers le monde. Autre exemple : le logiciel de correction de texte Criterion est conçu pour l'anglais américain standard ce qui le conduit à stigmatiser comme fautifs des usages rhétoriques et stylistiques pourtant avérés mais au sein de groupes ethniques ou sociaux non dominants (comme l'afro-américain)[32]. On voit comment, même au sein d'une langue donnée, les technologies de traitement automatique des langues peuvent restreindre la diversité linguistique.

---

31 Ashwin Ram *et al*., « Conversational AI : The science behind the Alexa Prize », *First Proceedings of the Alexa Prize*, 2017, p. 9

32 Nicky Hockly, « Automated writing evaluation », *ELT Journal*, vol. 73, n° 1, janvier 2019, p. 85. [En ligne] https://doi.org/10.1093/elt/ccy044 [consulté le 22 juin 2020].

Les indicateurs de performance jouent ici également un rôle clé : en plus de contribuer à la promotion des technologies de TAL auprès de futurs utilisateurs, ils orientent la recherche dans la mesure où les équipes cherchent à optimiser leurs algorithmes. Jusqu'à récemment, l'évaluation de la traduction automatique neuronale se faisait par comparaison avec la traduction humaine, le plus souvent à partir du modèle BLEU[33]. Or, depuis 2019, un autre indicateur est mis en avant par les plateformes : le taux d'engagement des utilisateurs. Au lieu de viser à rendre le plus fidèlement possible le texte source dans la langue cible, il s'agit de maximiser le nombre de clics ou le temps passé par l'internaute sur une page ou une interface donnée. Facebook a ainsi déposé un brevet pour un taux d'engagement[34] et Amazon cherche à développer la capacité conversationnelle d'Alexa, « mesurée en termes de durée, de tours de parole et de scores établis par des évaluateurs d'engagement[35] ». Toutefois, les indicateurs basés sur les taux d'engagement des utilisateurs pourraient aussi contribuer à un appauvrissement de la diversité linguistique : en effet les gens privilégient souvent soit des contenus très spécifiques, soit des informations controversées ou farfelues, soit les contenus étiquetés comme populaires (« les plus visionnés ») ou ceux qui confortent leur vision du monde (l'effet « bulle »)[36].

L'interaction humain/machine est une autre dimension de ces programmes qui mérite d'être explorée. La plupart des outils de TAL sont conçus pour permettre une interaction entre des acteurs techniques (un réfrigérateur intelligent, une plateforme de vente) et des acteurs humains. Pour que ces outils fonctionnent de manière fluide, il faut réduire les variations dans l'expression d'une idée donnée. Autrement dit, la requête émise par l'être humain doit être simplifiée. « Les requêtes peuvent être soit des *invites formelles* prédéfinies que l'utilisateur doit connaître pour déclencher l'action voulue, soit des requêtes exprimées dans la *langue naturelle*, soit *des données issues de capteurs* souvent récoltées à l'insu de l'utilisateur[37]. » Le

---

33 Joss Moorkens, Sheila Castilho, Federico Gaspari et Stephen Doherty (dir.), *Translation Quality Assessment. From Principles to Practice*, Cham, Springer International Publishing, p. 9 *sqq*.

34 Esther Bond, « Facebook patents alternative to "expensive" BLEU », *Slator*, 22 août 2019. [En ligne] https://slator.com/technology/facebook-patents-alternative-to-expensive-bleu/ [consulté le 21 novembre 2019].

35 Ashwin Ram *et al*., « Conversational AI : The science behind the Alexa Prize », *First Proceedings of the Alexa Prize*, 2017, p. 12.

36 Heather O'Brien, « Exploring user engagement in online news interaction », *Proceedings of the American Society for Information Science and Technology*, 11 janvier 2012. [En ligne] https://doi.org/10.1002/meet.2011.14504801088 [consulté le 3 juillet 2020].

37 Robin Knote, Andreas Janson, Matthias Söllner et Jan Marco Leimester, « Classifying smart personal assistants : An empirical cluster analysis », *Proceedings of the 52d Hawaii International Conference on System Sciences*, 2019, p. 2027.

bouton d'aide sur Siri affiche ainsi différentes invites préprogrammées comme « Montre-moi mes photos », « Programme une réunion à 9 h » ou « FaceTime avec John ». Google Translate et d'autres applications d'interprétation automatique telles que VoiceTra, mise au point au Japon en prévision des Jeux olympiques, exigent aussi que l'utilisateur simplifie et rationalise ses phrases en amont pour que la machine puisse les traiter correctement. Le traitement des requêtes contribue ainsi à la standardisation de la langue[38].

L'évolution récente du marché des technologies linguistiques limite également la diversité. L'un des sujets les plus préoccupants est l'arrivée sur le marché d'une poignée de géants du numérique, à savoir Google, Amazon, Facebook, Apple, Microsoft et leurs équivalents asiatiques. Ces acteurs dominent le marché à plusieurs titres : par leur taille, par les sommes colossales qu'ils investissent en TAL et en raison du nombre d'utilisateurs de leurs services. Les dernières projections prévoient que le nombre d'utilisateurs d'assistants personnels autonomes tels qu'Alexa (Amazon), Siri (Apple) ou Cortana (Microsoft) passera de 390 millions en 2015 à 1,8 milliard en 2021, soit une progression annuelle moyenne des ventes de 3 milliards de dollars[39]. Le marché des technologies de traitement des langues pourrait bientôt se transformer en oligopole à franges dominées par de très grosses entreprises américaines et chinoises dont le principal objectif n'est pas la mise à disposition de solutions linguistiques de qualité mais davantage d'étendre leur mainmise sur les interactions humaines en ligne.

Ces remarques nous conduisent à un dernier point concernant la diversité des langues : le type de gouvernance mis en œuvre par les principaux fournisseurs de technologies de traitement des langues. Tous s'efforcent d'intégrer des technologies de modération de contenu au sein de leurs outils. Amazon se vante d'avoir mis en place une politique proactive permettant de rendre l'expérience de l'utilisateur pleinement positive ; or dans un certain nombre de cas et vu sous un autre angle, cette dernière s'apparente tout simplement à de la censure.

> Alors que notre système de modération initial reposait sur un mécanisme de « liste noire », nous avons aussi réalisé des efforts dans le sens d'une classification plus fine et sensible au contexte pour identifier 1) les contenus vulgaires 2)les contenus sexuels 3) les contenus incitant à la haine raciale 4) d'autres formes d'incitation à la haine et 5) les contenus violents. Ce système sera intégré d'office aux futurs

38 Rory Smith, « The Google Translate world cup », *New York Times*, 13 juillet 2018. [En ligne] https://www.nytimes.com/2018/07/13/sports/world-cup/google-translate-app.html [consulté le 22 juin 2020].

39 Robin Knote, Andreas Janson, Matthias Söllner et Jan Marco Leimeister, « Classifying smart personal assistants : An empirical cluster analysis », *Proceedings of the 52d Hawaii International Conference on System Sciences*, 2019, p. 2024.

> concours Alexa Prize, étant essentiel à la garantie d'une expérience usager positive[40].

En dehors des règles de publication et de diffusion de contenu auto-édictées par les plateformes, il n'existe pas à l'heure actuelle d'organismes de contrôle, ni même d'accord sur ce que seraient les principes d'une bonne gouvernance linguistique en ligne.

Conclusion

L'effet des outils de traitement automatique des langues sur le paysage mondial des langues en ligne est complexe. Les algorithmes gratuits de traduction neuronale et les assistants personnels intelligents permettent à de nombreux utilisateurs à travers le monde de produire et de diffuser du contenu dans différentes langues, promouvant ce que j'appellerais un multilinguisme d'interface en matière de quantité, de visibilité et d'accessibilité. Pourtant, à y regarder de plus près, cette évolution se fait aux dépens de la diversité linguistique. Les niveaux de compétence linguistique, aussi bien en langues étrangères que dans la langue maternelle, peuvent pâtir de l'omniprésence de technologies linguistiques qui simplifient et standardisent l'expression. L'intelligibilité mutuelle est certes enrichie par la traduction automatique, mais cette dernière n'est pas conçue pour rendre compte des intraduisibles, de toutes les idiosyncrasies et les spécificités d'une communauté linguistique donnée. Enfin, le fait que les grandes plateformes et les réseaux sociaux puissent encadrer et contrôler la plupart de nos interactions en ligne, selon des principes de gouvernance qui leur sont propres et notamment ancrés dans une conception anglo-saxonne des échanges, est susceptible de réduire l'expression de la diversité des cultures. Si on admet que les langues sont bien plus que des vecteurs de communication, qu'elles constituent la pierre angulaire de nos pratiques et de nos identités culturelles, on peut s'inquiéter de voir une poignée d'acteurs économiques dominants œuvrer à la diffusion à grande échelle de nombreuses technologies automatiques de la langue. À l'instar des technologies d'IA dans d'autres secteurs, ces programmes devraient être soumis à des audits et les prestataires s'engager à rendre des comptes. Il serait temps de réunir les plateformes, les institutions (éducatives, culturelles, juridiques) et les communautés d'utilisateurs et pour engager un vrai débat sur le rôle et les conséquences du traitement automatique des langues dans nos sociétés.

---

40 Ashwin Ram *et al*., « Conversational AI : The science behind the Alexa Prize », *First Proceedings of the Alexa Prize*, 2017, p. 6.

Bibliographie

Bond, Esther, « What's so massive about Google's massively multilingual neural machine translation? », *Slator*, 18 juillet 2019. [En ligne] https://slator.com/technology/whats-so-massive-about-googles-massively-multilingual-neural-machine-translation/ [consulté le 18 septembre 2019].

Bond, Esther, « Facebook patents alternative to "expensive" BLEU », *Slator*, 22 août 2019. [En ligne] https://slator.com/technology/facebook-patents-alternative-to-expensive-bleu/ [consulté le 21 novembre 2019].

Charlton, Emma, « The Internet has a language diversity problem », Forum économique mondial, 13 décembre 2018. [En ligne] https://www.weforum.org/agenda/2018/12/chart-of-the-day-the-internet-has-a-language-diversity-problem/ [consulté le 18 septembre 2019].

Citton, Yves, *Médiarchies*, Paris, Seuil, 2017.

Diño, Gino, « Facebook says it now has a powerful tool to think about language problems in a language agnostic way », *Slator*, 6 mai 2019. [En ligne] https://slator.com/technology/at-f8-facebook-says-it-now-has-a-powerful-tool-to-think-about-language-problems-in-a-language-agnostic-way [consulté le 18 septembre 2019].

Edmond, Charlotte, « This is when a robot is going to take your job, according to Oxford University », 26 juillet 2017. [En ligne] https://www.weforum.org/agenda/2017/07/how-long-before-a-robot-takes-your-job-here-s-when-ai-experts-think-it-will-happen/ [consulté le 18 septembre 2019].

Haberkorn, B. J., « Amazon Polly voices in Alexa skills now generally available », 23 octobre 2018. [En ligne] https://developer.amazon.com/fr/blogs/alexa/post/baee53c1-5b03-4580-b57a-ee9510413354/amazon-polly-voices-in-alexa-skills-now-generally-available [consulté le 17 décembre 2019].

Hockly, Nicky, « Automated writing evaluation », *ELT Journal*, vol. 73, n° 1, Janvier 2019. [En ligne] https://doi.org/10.1093/elt/ccy044 [consulté le 22 juin 2020].

Internet retailing, « Zara. Communication without borders », 7 août 2017. [En ligne] https://internetretailing.net/research-articles/research-articles/zara-communication-without-borders [consulté le 17 décembre 2019].

Johnson, Khari, « Facebook Messenger launches translations by intelligent assistant M », 1er mai 2018. [En ligne] https://venturebeat.com/2018/05/01/facebook-messenger-launches-translations-by-intelligent-assistant-m/ [consulté le 18 septembre 2019].

Kemp, Charlotte, « Defining multilingualism », in Aronin, Larissa et Hufeigein, Britta (dir.), *The Exploration of Multilingualism*, Amsterdam, John Benjamins Publishing, 2009.

Kenny, Dorothy, « Machine translation », in Rawling, Piers et Wilson, Philip (dir.), *The Routledge Handbook of Translation and Philosophy*, Abingdon, Routledge, 2018, p. 428-445.

Knote, Robin, Andreas Janson, Matthias Söllner et Jan Marco Leimeister, « Classifying smart personal assistants : An empirical cluster analysis », *Proceedings of the 52d Hawaii International Conference on System Sciences*, 2019.

Larsonneur, Claire, « The disruptions of neural machine translation », *Spheres*, n° 5, novembre 2019. [En ligne] http://spheres-journal.org/the-disruptions-of-neural-machine-translation/ [consulté le 17 décembre 2019].

Leppänen, Sirpa et Peuronen, Saija, « Multilingualism on the Internet », in Martin-Jones, Marilyn, Blackledge, Adrian et Cresse, Angela (dir.) *The Routledge Handbook of Multilingualism*, Londres/New York, Routledge, 2015, p. 384-403.

Mazareanu, Elena, « Market size of the global language services market », 9 août 2019. [En ligne] https://www.statista.com/statistics/257656/size-of-the-global-language-services-market/ [consulté le 3 juillet 2020].

Doherty, Stephen (dir.), *Translation Quality Assessment. From Principles to Practice*, Cham, Springer International Publishing, 2018.

Pajot, Philippe, « La traduction automatique s'attaque aux langues rares », *La Recherche*, n° 554, décembre 2019, p. 20 *sq*.

O'Brien, Heather, « Exploring user engagement in online news interaction », *Proceedings of the American Society for Information Science and Technology*, 11 janvier 2012. [En ligne] https://doi.org/10.1002/meet.2011.14504801088 [consulté le 3 juillet 2020].

Paolillo, John, « Language diversity on the Internet : Examining linguistic bias », in *Measuring Linguistic Diversity on the Internet* (UNESCO Institute for Statistics), Paris, UNESCO, 2005.

Pasquinelli, Matteo, « How a machine learns and fails – A grammar of error for artificial intelligence», *Spheres*, n° 5, novembre 2019. [En ligne] http://spheres-journal.org/how-a-machine-learns-and-fails-a-grammar-of-error-for-artificial-intelligence/ [consulté le 13 décembre 2019].

Ram, Ashwin *et al.*, « Conversational AI : The science behind the Alexa Prize », *First Proceedings of the Alexa Prize*, 2017.

Shaji, Mishel, « What is the most translated website in the world? », 17 mars 2019. [En ligne] https://www.quora.com/What-are-the-top-three-most-translated-Websites-and-how-many-languages-do-they-each-accommodate [consulté le 18 septembre 2019].

Smith, Rory, « The Google Translate World Cup », *New York Times*, 13 juillet 2018. [En ligne] https://www.nytimes.com/2018/07/13/sports/world-cup/google-translate-app.html [consulté le 22 juin 2020].

Technavio, « Globalization of business to boost growth ». [En ligne] https://www.businesswire.com/news/home/20200529005034/en/Machine-Translation-Market-2020-2024-Globalization-Business-Boost [consulté le 3 juillet 2020].

Villani, Cédric, *For a Meaningful Artificial Intelligence. Towards a French and European Strategy*, mars 2018. [En ligne] https://www.aiforhumanity.fr/en/ [consulté le 20 septembre 2019].

Wiggers, Kyle, « Facebook founds AI Language Research Consortium to solve challenges in natural language processing », 28 août 2019. [En ligne] https://venturebeat.com/2019/08/28/facebook-founds-ai-language-research-consortium-to-solve-challenges-in-natural-language-processing/ [consulté le 17 décembre 2019].