LANGAGES

APPEL À CONTRIBUTIONS POUR LE N° 237 (1/2025)
Date limite de soumission des propositions : 15 novembre 2023

Titre provisoire : *Intelligence artificielle, analyse de corpus et diversité linguistique*

Coordonné par Rachele Raus (Université de Bologne, Italie) et Michela Tonti (Université de Bergame, Italie)

Les évolutions récentes dans l'« apprentissage profond » (*deep learning*) par lequel un dispositif basé sur des réseaux neuronaux imitant ceux du cerveau humain apprend par le biais des données (Le Cun 2019), ont permis à l'intelligence artificielle de connaître un essor considérable dans tous les domaines des activités humaines et de la recherche (voir, entre autres, Zouinar 2020 ; von Braun *et alii* 2021). Dans le domaine linguistique (Tavosanis 2018), l'IA a trouvé son application majeure dans l'industrie des langues à des fins diverses, comme, par exemple, la traduction (Kohen 2020) ou l'apprentissage des langues (Miras *et alii* 2019).

Toutefois, cette évolution soulève de nombreuses questions concernant, entre autres, l'impact de l'anglais en tant que langue pivot pour l'apprentissage profond des dispositifs linguistiques (Kim *et alii* 2019 ; Vetere 2023), – comme ceux d'écriture, d'interprétation, de traduction, de doublage, etc. – ou bien les implications sociales, culturelles et éducatives liées à l'utilisation de données massives (*big data*), c'est-à-dire de vastes corpus multilingues (Caliskan *et alii* 2017 ; Rastier 2021).

En effet, les études « critiques » de l'intelligence artificielle portent surtout sur le manque de ressources informatiques dans certaines langues, qui seraient donc mises en péril par l'utilisation de l'IA dans l'industrie des langues (Moorkens 2022, voir aussi les rapports du projet *European Language Equality*[1]). Mais ces études portent aussi sur la standardisation des langues aux dépens du multilinguisme (Larsonneur 2021 ; Raus *et alii* 2023), ainsi que sur les biais introduits par l'apprentissage non supervisé, c'est-à-dire sans intervention humaine. Dans ce cas, en effet, la machine apprend via des données biaisées et, par conséquent, génère des discriminations véhiculées par le langage (Bartoletti 2020 ; Stypinska 2021 ; pour les discriminations liées au « genre », voir, entre autres, Marzi 2021 ; Savoldi *et alii* 2021).

Ces questions découlent des approches et des méthodes utilisées dans la recherche sur l'IA. En effet, les approches statistiques d'apprentissage profond fondées sur des corpus, qui sont les plus répandues actuellement (Chaumartin 2020), sont également les plus problématiques, notamment lorsqu'elles se passent de l'intervention humaine (comme il arrive, le plus souvent, dans le cas de ChatGPT) alors que cette dernière permet justement d'atténuer les questions liées aux biais ou à la standardisation des langues (Attanasio *et alii* 2021 ; Raus *et alii* 2022). La possibilité d'utiliser éventuellement le système du tranfert positif (Artetxe, Schwenk 2019), par lequel on transfère à d'autres langues des systèmes ou des résultats obtenus pour une seule langue qui normalement est l'anglais, et la tentative de découvrir des représentations universelles des langues par ce

[1] https://european-language-equality.eu/deliverables/ (voir D1.1, D1.3, D1.4-36 et D1.38-39).

type de traitement automatique suscitent elles aussi de nombreuses questions, notamment à l'égard de l'universalité réelle des significations extraites (Yvon, à paraître).
Partant d'une approche alternative et de tâches différentes, des chercheurs de l'ETIS (Pitti *et alii* 2021) se sont inspirés de l'acquisition de la parole par l'enfant pour apprendre à une IA à parler. Ce faisant, ils ont réussi à s'affranchir des quantités volumineuses de données utilisées généralement pour l'apprentissage des réseaux neuronaux. De la même manière, une équipe de recherche au MIT (Ross *et alii* 2018) a permis à un dispositif d'IA d'apprendre à interpréter le langage par l'observation de son propre environnement, sans passer donc par l'apprentissage de données préalables.

Toutes ces approches enrichissent le débat sur l'IA et sur son utilisation dans le domaine des langues, en favorisant également le dépassement des disciplines traditionnelles et la création de nouveaux profils professionnels à la croisée entre sciences humaines et computationnelles (Ferraresi *et alii* 2021), ce qui ouvre de nouvelles perspectives de recherche transdisciplinaires.

Ces études nous amènent donc à réfléchir sur l'IA : d'une part, par rapport au corpus, dans le cas des approches statistiques, et de l'autre, en matière de diversité et d'évolutions des langues en général.

Argumentaire

Dans le cadre du contexte esquissé, plusieurs thématiques peuvent être abordées pour ce qui est des approches de l'IA en relation aux langues, notamment :

- Les problèmes morphosyntaxiques liés au traitement automatique des langues par l'IA qui se répercutent sur les langues et sur leurs évolutions (formes grammaticales ou syntaxiques incorrectes) comme, par exemple, l'accord des noms de métiers et de professions au féminin qui disparaît dans la traduction automatique entre langues romanes provoquant la perte de l'accord ;
- Les formes de standardisation lexicale — tels, entre autres, les « emprunts de sens » (Paquet-Gauthier 2018) ou plus généralement la « convergence lexicale » (Hermand 2014) —, provoquées par l'utilisation de l'IA dans le domaine linguistique, ce qui entraîne l'uniformisation du sens des mots des différentes langues, voire leur harmonisation graphique ;
- Les conséquences que l'utilisation des corpus multilingues pour l'apprentissage profond de l'IA entraîne sur la diatopie linguistique, par exemple lorsque la machine apprend à partir de corpus rédigés dans des langues véhiculaires et/ou côntrolées, comme celles qui sont utilisées par les organisations internationales (telles les langues officielles des Nations unies, celles de l'Union européennes, etc.), et finit par générer du texte qui ne respecte pas les spécificités des langues nationales équivalentes ;
- Les répercussions éventuelles que la traduction automatique basée sur l'apprentissage profond et donc sur l'IA ont sur la diatopie linguistique, par exemple quand c'est l'anglais qui est utilisé comme langue pivot ;
- Les retentissements du modèle informatique plus ou moins supervisé, c'est-à-dire plus ou moins surveillé par l'humain, sur les langues dans les dispositifs linguistiques pour ce qui est de la perte ou vice-versa de la valorisation de la diatopie linguistique ;

- Les conséquences du choix de l'unité de base de l'apprentissage profond (unité sous-lexicale, lexicale ou « passage », selon Rastier 2007) sur le traitement automatique des langues, chaque unité permettant au dispositif d'être plus ou moins performant et de préserver ou pas la spécificité d'une langue, par exemple en permettant d'en repérer les caractéristiques intrinsèques à des fins d'analyse mais aussi de traitement ultérieur comme, entre autres, dans le cas de la traduction.

Sans prétendre à l'exhaustivité, ce ne sont là que quelques-uns des sujets que ce numéro entend aborder pour répondre à des questions précises sur l'IA et sur l'évolution des langues et de la diversité linguistique.

Objectifs du numéro

Ce numéro de *Langages* veut être l'occasion de dresser un bilan et d'encourager une réflexion commune sur l'intelligence artificielle par rapport aux corpus et aux langues afin de répondre aux questions que les nouvelles technologies fondées sur l'apprentissage profond soulèvent.

D'abord, il s'agit de voir comment la modélisation informatique et l'utilisation de l'IA se répercutent sur les langues et s'il existe des approches qui, tout en utilisant (ou pas) des corpus, peuvent s'avérer meilleures quant au respect de la diatopie linguistique.

Le deuxième objectif de ce numéro est d'analyser les conséquences de l'utilisation de l'IA sur la morphologie, le lexique et la syntaxe des langues, en considérant également — le cas échéant — les aspects sémantiques. Il s'agit aussi de vérifier si et de quelle manière les langues véhiculaires et/ou côntrolées utilisées notamment à l'échelle internationale, comme par exemple celles des organisations internationales, facilitent la standardisation lexicale lors de l'apprentissage profond basé sur des corpus qui sont rédigés dans ces langues.

Enfin, à travers l'étude de différents cas, issus de la traduction automatique neuronale, de la génération automatique du texte, de la reconnaissance automatique de la parole, etc., en d'autres termes du traitement automatique des langues basé plus généralement sur l'apprentissage profond (Allauzen, Schütze 2018), il s'agit d'envisager les apports et aussi les répercussions que l'intelligence artificielle aura sur les langues et sur leurs évolutions.

Références

Allauzen, Alexandre, Schütze, Hinrich (eds), 2018, « Apprentissage profond pour le traitement automatique des langues », *Apprentissage automatique des langues* Vol. 59, n°2. https://www.atala.org/content/apprentissage-profond-pour-le-traitement-automatique-des-langues

Artetxe, Mikel, Schwenk, Holger, 2019, « Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond », *Transactions of the Association for Computational Linguistics* n° 7 : 597-610.

Attanasio, G., Greco, S., La Quatra, M., Cagliero, L., Tonti, M., Cerquitelli, T. & Raus, R., 2021, « E-MIMIC : Empowering Multilingual Inclusive Communication », *2021 IEEE International Conference on Big Data*. https://bigdataieee.org/BigData2021/

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

Braun, J. (von), Archer, M. S., Reichberg, G. M., Sánchez Sorondo, M., 2021, *Robotics, AI, and Humanity. Science, Ethics, and Politics*, Suisse : Springer.

Caliskan, A., Bryson, J.J. & Narayan, A., 2017, « Semantics derived automatically from language corpora contain human-like biases », *Sciences* n°356(6334) : 183-186.

Chaumartin, François-Régis & Lemberger, Pirmin, 2020, *Le traitement automatique des langues. Comprendre les textes grâce à l'intelligence artificielle*. Malakoff : Dunod.

Ferraresi, A., Aragrande, G., Barrón-Cedeño, A., Bernardini, S. & Miličević Petrović, M., 2021, *Competences, skills and tasks in today's jobs for linguists: Evidence from a corpus of job advertisements. UPSKILLS Intellectual output 1.3*. https://zenodo.org/record/5030879

Hermand, Marie-Hélène, 2014,. «Le discours eurorégional. Indices convergents de légitimation d'un espace institutionnel», *Mots. Les langages du politique*, n° 106 : 71-86.

Kim, Y., Petrov, P., Petrushkov, P., Khadivi, S. & Ney, H., 2019, « Pivot-based Transfer Learning for Neural Machine Translation between Non-English Languages », *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, November 3–7 : 866-876.

Kohen, Philipp, 2020, *Neural Machine Translation*, Cambridge : Cambridge University Press.

Larsonneur, Claire, 2021, « Intelligence artificielle et/ou diversité linguistique : le paradoxe du traitement automatique des langues », *Hybrid* n°7. https://journals.openedition.org/hybrid/650

Le Cun, Yann, 2019, *Quand la machine apprend. La révolution des neurones artificiels et de l'apprentissage profond*, Paris : Odile Jacob.

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

Moorkens, Joss, 2022, *Ethics and machine translation*. In Kenny Dorothy (ed), *Machine Translation for everyone. Empowering users in the age of artificial intelligence*. Berlin : Language Science Press : 121-140.

Paquet-Gauthier, Myriam, 2018, « Changements sémantiques sous l'influence de l'anglais : le cas de quatre 'emprunts de sens' en français au Québec (1992–2012) », dans C. Jacquet-Pfau, A. Napieralski et J.-F. Sablayrolles (eds), *Emprunts néologiques et équivalents autochtones : études interlangues*, Łódź, WUŁ : 201-228.

Pitti, A., Quoy, M., Boucenna, S. & Lavandier, C, 2018, « Brain-inspired model for early vocal learning and correspondence matching using free-energy optimization ». *PloS Computational Biology*, 2021.

Rastier, François, 2007, « Passages », *Corpus* n° 6. https://journals.openedition.org/corpus/832

Rastier, François, 2021, « Data vs Corpora », in D. Mayaffre & L. Vanni (eds) *L'intelligence artificielle des textes : des algorithmes à l'interprétation*, Paris : Champion, 203-245.

Raus, R., Tonti, M., Cerquitelli, T., Cagliero, L., Attanasio, G., La Quatra, M. & Greco S., 2022, « L'analyse du discours et l'intelligence artificielle pour réaliser une écriture inclusive », *Congrès mondial de Linguistique française 2022*. https://doi.org/10.1051/shsconf/202213801007

Raus, Rachele, Humbley, John, Silletti, Alida Maria & Zollo, Silvia (eds), 2023, *Multilingualism and language varieties in Europe in the age of artificial intelligence,* De Europa Special issue 2022, Milan : Ledizioni.

Ross, C., Barbu, A., Berzak, Y., Myanganbayar, B. & Katz, B., 2018, « Grounding language acquisition by training semantic parsers using captioned videos », *2018 Conference on Empirical Methods in Natural Language Processing*, Bruxelles, Association for Computational Linguistics : 2647–2656. https://aclanthology.org/D18-1285/

Savoldi B., Gaido M., Bentivoglio L., Negri M. & Turchi M., 2021, « Gender Bias in Machine Translation », *Transactions of the Association for Computational Linguistics* n° 9 : 845-874. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00401/106991/Gender-Bias-in-Machine-Translation

Stypinska, Justina, 2021, « AI ageism : new forms of ages biases and age discrimination in the era of algorithms and artificial intelligence », *Conference: Proceedings of the 1st International Conference on AI for People: Towards Sustainable AI*, CAIP 2021, 20-24 November 2021, Bologna, Italy. https://eudl.eu/doi/10.4108/eai.20-11-2021.2314200

Tavosanis, Mirko, 2018, *Lingue e intelligenza artificiale*, Rome : Carocci.

Vetere, Guido, 2023, « Elaborazione automatica dei linguaggi diversi dall'inglese : introduzione, stato dell'arte e prospettive », *De Europa Special issue 2022*, Milan : Ledizioni : 69-87.

Yvon, François, à paraître, « La traduction multilingue : analyse d'une prouesse technologique », *mediAzioni. Rivista online di studi interdisciplinari in lingue e culture*, Numero speciale : *L'intelligenza artificiale per la traduzione : verso una nuova progettazione didattica ?*, I. Cennamo, L. Cinato, M.M. Mattioda, A. Molino (eds).

Zouinar, Moustafa, 2020, « Evolutions de l'Intelligence artificielle : quels enjeux pour l'activité humaine et la relation Humain-Machine au travail », *Activités* n°17-1. https://journals.openedition.org/activites/4941#ftn1

Calendrier

- envoi des propositions avant le **: 15 novembre 2023**

A adresser à
rachele.raus@unibo.it
michela.tonti@unibg.it
cschnede@unistra.fr

Format des propositions : 1 à 4 pages présentant le propos, l'originalité de la contribution, la méthodologie adoptée le cas échéant, les résultats escomptés et les retombées ainsi qu'une bibliographie.

- accusé d'acceptation ou de refus de la proposition : avant le 10 janvier 2024

- envoi des articles : avant le 30 juin 2024

- envoi des évaluations : avant le 15 octobre 2024

- envoi des articles définitifs : avant le 15 novembre 2024

- publication du numéro : comme n°1 de 2025, donc en mars 2025.