Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

Inclusione ed elaborazione del linguaggio naturale nell'era dell'intelligenza artificiale generativa

AA. VV.

Politecnico di Torino

EMIMIC

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

La pubblicazione è finanziata su fondi del progetto PRIN 2022 PNRR
"Empowering Multilingual Inclusive Communication"
Finanziato dall'Unione Europea NextGenerationEU
PNRR - Missione 4 Istruzione e Ricerca
Componente 2 dalla Ricerca d'Impresa
Investimento 1.1., codice 2022WEFCFP - CUP J53D23007230006

E-MIMIC

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Ledizioni
The Innovative LEDIpublishing Company

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

ISBN ebook: 9791256004133

Graphics and page layout
Silvio Ortolani, SISHO - fotografia & archivi

Inclusione ed elaborazione del linguaggio naturale nell'era dell'intelligenza artificiale generativa

Inclusione ed elaborazione del linguaggio naturale nell'era dell'intelligenza artificiale generativa

I. Inclusione, elaborazione del linguaggio naturale e intelligenza artificiale

II. Il progetto E-MIMIC

III. Riflessioni a *latere*

"The limits of LLMs are the limits of the world": considerations on the role of linguistics in enhancing AI-generated language quality

Valeria Zotti, valeria.zotti@unibo.it
Università di Bologna

Introduction

In this short article I will present some observations that arose from the activities I carried out in 2024 as part of the PRIN 2022 E-MIMIC project.

In the first part, I will briefly describe the increasing focus on inclusive communication and the risks associated with the use of artificial intelligence systems, such as the loss of language diversity and the spread of biases, which I will illustrate with a few examples from neural machine translation.

In the second part, I will discuss how the flattening of language and the loss of multilingualism caused by AI systems may coincide with the loss of the ability to conceptualize the world differently from dominant models. Observation of language acquisition and development in children reveals the social conditioning that permeates a language and the role of social experience in the comprehension process, a factor that distinguishes human competence from artificial language production.

Linguistics proves to be of central importance in understanding the complexity of human language, as well as the need for strong human involvement in data-driven methods, as is being done in the E-MIMIC project whose innovative deep-learning methodology aims to develop artificial intelligence systems that can serve to promote equality in communication.

1. Deep Learning & Gender Bias in Language and in AI

Awareness of the risks that the development of Artificial Intelligence, particularly Generative AI (GenAI), may cause for the preservation of linguistic and cultural diversity and the protection of gender equality and minority rights has grown significantly in the scientific community over the past few years (Raus *et al.*, 2022; Raus, 2024; Cennamo *et al.*, 2023; Mattioda, 2024).

The rapid and unprecedented progress of artificial intelligence in all areas of human endeavor, and particularly a branch known as Deep Learning, which enables neural network-based algorithms to learn "autonomously" on large amounts of data (Le Cun, 2019), has raised serious questions, both in academia and in industry and institutions, about the provenance and quality of data used to train these systems.

As noted by New York Times tech journalist Cade Metz, in the 2010s, "As Google and other tech giants adopted the technology, no one quite realized it was learning the biases of the researchers who built it. These researchers were mostly white men, and they didn't see the breadth of the problem until a new wave of researchers — both women and people of color — put a finger on it" (Metz, 2021).

The lack of diversity and inclusion that characterizes the design of AI systems is thus a major concern today. In the field of text generation, AI systems reproduce what humans say and write. If a system is fed on human biases (conscious or unconscious), it will produce results full of these same biases. In the current Large Language Models on which AI systems are trained, the data are heterogeneous and not always traceable. These data risk reinforcing discrimination and bias by giving them the appearance of objectivity.

To illustrate how these systems risk amplifying stereotypes, I will give a first example related to the massive spread of bias in neural machine translation, which, as Yvon (2023: A19) explains, is mainly based on machine learning methods exploiting large corpora of parallel texts aligned at sentence level.

In a second example I will briefly show that another disadvantage of generative AI is that, because of the dominance of training data in English, the preservation of multilingualism is endangered.

1.1 Bias in Neural Machine Translation systems

The following example shows "gender bias" present in the data generated by AI systems taken from a small test performed in 2021 on the well-known neural machine translation software, DeepL.

While in Italian it is possible to omit the subject personal pronoun preceding the verb, resulting in neutral sentences that do not indicate whether the subject is masculine or feminine, in English and French it must be specified.

The output of this NMT system in both English and French (Tab. 1) suggests that numerous stereotypes were present in the data used to train it, made explicit by the use of the female personal pronoun for roles traditionally delegated to the female sex, such as childcare. Other experiments have shown that if the subject personal pronoun is not specified, DeepL generally has a preference for the masculine.

Traduci da Italiano	***Traduci in Inglese (US)***
Pratica sport. Si occupa dei bambini. Lavora. Guida l'auto.	He plays sport. She takes care of children. He works. He drives a car.
Traduci da Italiano	***Traduci in Francese***
Pratica sport. Si occupa dei bambini. Lavora. Guida l'auto.	Il fait du sport. Elle s'occupe des enfants. Il travaille. Il conduit une voiture.

Tab. 1. Test DeepL 2021

Three years later, when the same test was repeated in 2024 the same stereotypes were generated again. Activities such as playing sports, working and driving a car are still presented by the British and American English translations as masculine activities, while, in the sentence related to childcare, the software hesitates and omits the subject pronoun thus returning an incorrect sentence (Tab. 2). In the French translation, on the other hand, childcare is attributed to a masculine subject (Fr. pronoun "il" = En. "he"), suggesting that the French data may be less "biased" than the English data.

Traduci da Italiano	***Traduci in Inglese britannico***
Pratica sport. Si occupa dei bambini. Lavora. Guida l'auto.	He plays sport. Takes care of children. Works. He drives a car.
Traduci da Italiano	*Traduci in Inglese americano*
Pratica sport. Si occupa dei bambini. Lavora. Guida l'auto.	Plays sport. Takes care of children. He works. He drives a car.
Traduci da Italiano	*Traduci in Francese*
Pratica sport. Si occupa dei bambini. Lavora. Guida l'auto.	Il pratique le sport. Il s'occupe des enfants. Il travaille. Il conduit une voiture.

Tab. 2. Test DeepL 2024

This simple test raises further questions about the preponderance of training data in English as a potential source of bias.

1.2 English language dominance in Human and Machine Translation

Given that these technologies rely above all on data in English, a language which in many cases is not gender-specific, their ability to generate inclusive texts in other languages (e.g. in the Romance languages involved in the E-MIMIC project), is rather limited.

English is dominant not only for economic reasons, with leading research on AI taking place largely in English-speaking countries, but also for linguistic reasons. Morphologically complex languages (such as Italian, French, and Spanish) are much more difficult to handle than English, while English has the advantage of being morphologically simple (e.g. it has few verb inflections, no gender markings), so creating IT language models for other languages requires more computer engineering work and is more expensive (Vetere, 2023).

An example of the impoverishment of lesser-represented languages due to this sort of "colonization" of English is provided by a study by Henkel (2024) focusing on a comparison of human translation and machine translation, which shows that machine translations have reduced lexical variety. The example reported in this study is about the verb "think" in English which, in the case of human translation, is translated in French in 29% of cases by "*penser*" and in 28% by "*croire*" and for the remaining 40% by other lexemes or expressions, while DeepL translates in 74% of cases with "*penser*" and a few other variants. As noted by Henkel (2024), machine translation applies a limited number of "recipes": it tends to favour one single translation strategy with a small number of alternative solutions.

The real risk of NMT-induced language impoverishment is thus astoundingly obvious in the statistical data, the dangers of which cannot be overstated. Indeed, linguistic diversity is an expression of diversity of thought, but also of human diversity with respect to gender and ethnic or age differences, disabilities, etc. As follows from Ludvig Wittgenstein's famous aphorism "the limits of language are the limits of my world", if human language becomes impoverished, under the influence of machine-generated language, the risk in a not-too-dystopian future of an impoverishment of human thought is real, a recurring theme in science fic-

tion[1] literature and film. In other words, with this flattening of language generated by AI systems, the loss of multilingualism may coincide with the loss of the ability to conceptualize the world differently from dominant models. This ability is innate in human beings but the social conditioning we undergo from the moment we acquire a language affects our ability to conceptualize reality differently.

2. Social conditioning in Human and AI generated language

An example taken from my personal experience of observing how children acquire language reveals the extent to which language is a social product and is imbued with unconsciously transmitted stereotypes.

When my daughter was attending kindergarten at the age of 4 and could not yet read or write, I informed her one day that she would be attending a party the next day with 'the children' (in Italian, "*i bambini*") from her class. She stared hard at me and replied, "Mom, why do you say '*I bambinI*' (masculine in Italian) and not '*lE bambinE*' (feminine in Italian)? In my class there are more girls than boys!".

I explained to her in simple words that in Italian we use the generic masculine to refer to both sexes. After protesting indignantly that "it's not fair," she suggested "*bambin**" as a new form that we would use in our family to indicate that we were all equal. Having begun at that time to take an interest in inclusive language, I recognized in this attempt to eliminate sexism in the Italian language the motive for the well-known *Recommendations* (Sabatini, 1987) and the many guidelines that have been published in the last 50 years to promote gender-neutral and non-discriminatory use of language.

This brief exchange with my daughter resonated in me as it made me more aware of how much social conditioning we undergo through language acquisition. Children are born pure with an innate sensitivity that is gradually lost as they learn a language.

[1] As observed during the *ALMA AI Film Festival. Pre-Visions of Artificial Intelligence*, sponsored by the ALMA AI research center and organized by Milano Michela and Zotti Valeria in spring 2024 at the *Cineteca* di Bologna.

My daughter naturally perceived an injustice conveyed by language, namely the occultation of women, which is the reflection of centuries of women's inferior social status. As observed by the father of linguistics, Ferdinand de Saussure, "language is social", it is a social product of the language faculty and a collective reality, a set of conventions that the community adopts (Saussure, 1916).

Human beings, as they learn a language also integrate stereotypes conveyed by language that are socially induced, the result of language habits and modes of interaction repeated over time within society. Over time, a child who becomes an adult loses the innate ability to recognize these stereotypes and unconsciously integrates social injustice.

My daughter's example shows another key factor, the fundamental role of "experience." She was aware that three-quarters of her class were females, and this experience, which was also social in nature, enabled her to grasp on a cognitive level the contradiction that exists between reality and the fictitious representation of reality conveyed by language.

Unlike children, AI systems have access to unimaginably large quantities of data but do not experience it socially in the form of real interactions with other living beings within a community (probably one of the reasons why they have no sense of humour yet, as demonstrated by Veale, 2021). Although there is research on "Embodied AI", which is generative IA inserted in a robotic body which should improve its performance, machines are a long way from social and pragmatic language competency (e.g. knowing what to say in the correct situation). This leads us to the concept of discursive diversity of which IT developers must become aware. Linguistics can be very helpful to raise this awareness.

3. Discursive diversity and the role of linguistics

Linguistics is the scientific study of human language in all its manifestations and structures. Linguists are fully aware of the essential notion of "discourse type". Language is not monolithic and is characterized by the co-existence of many "discourse types" and "registers," with different, and sometimes even contradictory or incompatible characteristics. The problem with LLMs is that they contain

raw data taken largely from the Internet (Naik *et al.*, 2024) in which some discourse types are under-represented, others are over-represented, and all are amalgamated as if they were the same type. These data cannot be considered as representative of human language.

The question of "representativeness" has been debated for many years in the field of corpus linguistics, well before the advent of AI. Chomsky (1957: 15) rejected corpus linguistics because he considered that corpus data would always be insufficient to describe language competency or the language system as a whole. Even a text corpus of several billion words is only a small sample compared to the amount of language that hundreds of millions of speakers produce in one day, on and off the internet. We can only comprehend language in its diversity and complexity and obtain higher-quality results from AI, in text generation or machine translation, by distinguishing different discourse types one at time, using specialized datasets.

The experience of the E-MIMIC Project is innovative in this respect insofar as the material used to train the neural networks is authentic and the analysis and preparation of the data was carried out by linguists following linguistic-discursive criteria designed to ensure inclusiveness (Raus *et al.*, 2022).

The role of linguistics is thus fundamental in IT research in order to understand the fundamental characteristics of human language and the reasons for which we recognize AI-generated language as "artificial". These are hotly debated issues among those involved in natural language generation and machine translation.

A fascinating book entitled *Intelligence artificielle, intelligence humaine: la double énigme* (The Double Enigma: Artificial Intelligence and Human Intelligence), by French philosopher and mathematician Daniel Andler (2023), offers a compelling enquiry into what "human" or "natural" intelligence is and the various difficulties that AI has encountered. Indeed, since its inception in the 1960s, AI has been claimed to have the potential to equal human intelligence, yet this initial ambition remains an illusion.

As far as language is concerned, generative AI has made astonishing progress in emulating human beings, but machines are still a long way from 'autonomy' with respect to humans because they lack true understanding of the complexity of human language, and they lack real-world human experience. Thus, linguistics is essential to

take into account the role that humans play as the era of AI unfolds. If AI lowers the quality of the language humans speak, if humans allow language to be impoverished and if people become accustomed to communicating and interacting in machine-like ways, the risk that we might lose the ability to distinguish between human and artificial intelligence and forget our humanity is real.

Conclusion

Political actions that have been conducted over the past few years have been aimed at re-educating speakers of a language so that they adopt ostensibly "new" but actually "innate" communicative practices that give recognition to all individuals. Some initiatives, such as that of the president of the University of Trent who, in March 2024, published University Regulations written using the 'overextended feminine' for offices and gender references (e.g., "*decana*" "dean-fem." "*rettrice*" "rector-fem." "*segretaria*" "secretary-fem." even for men), is, as stated by the president himself, "A symbolic act to demonstrate equality starting with the language of our documents." Pr. Flavio de Florian said that he felt excluded from this all-female document and understood the necessity of promoting neutral and non-sexist language.

It is paradoxical that in the same era in which we are becoming aware of the pervasiveness of sexist stereotypes in language, generative AI systems "amplify" these biases. Thus, it is essential to develop a linguistic data policy, to make computer scientists aware of this need, of the fact that methods used to collect data are crucial, that *"le manque de données textuelles de qualité entraîne une sous-représentation de certaines langues-cultures"*[2] (Mattioda, 2024: 5), and that relying on linguistic criteria is efficient. The E-MIMIC project and other commendable initiatives, such as the national AI research network FAIR, have shown how linguistics can help improve the quality of language output from AI and avoid the kind of linguistic impoverishment and biases that have so far plagued AI-generated language.

[2] "the lack of quality textual data leads to an under-representation of certain language-cultures" (Mattioda 2025: 5).

References

Andler Daniel (2023). *Intelligence artificielle, intelligence humaine: la double énigme*. Paris: Gallimard.

Cennamo Ilaria, Cinato Lucia, Mattioda Maria Margherita, Molino Alessandra (2023). "Promoting multilingualism and inclusiveness in educational settings in the age of AI." In: Tasa Fuster Vicenta, Monzó-Nebot Esther and Rafael Castelló-Cogollos (eds). *Repurposing language rights. Guiding the uses of artificial intelligence*. València: Tirant lo Blanch, 277-315.

Chomsky Noam (1957). *Syntactic Structures*. La Haye: Mouton.

Henkel Daniel (2024). "Verbs of cognition in translation between English and French". In: Dister Anne and Longrée Dominique (eds). *JADT 2024 Mots comptés, textes déchiffrés*. Louvain-La-Neuve: Presses Universitaires de Louvain, 399-408.

Le Cun Yann (2019). "Deep Learning Hardware: Past, Present, and Future". In: *2019 IEEE International Solid-State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 12-19.

Mattioda Maria Margherita (2024). "Intelligence artificielle et diversité linguistique. Quelle gestion équitable pour la garantie des droits linguistiques ?". *Language Problems and Language Planning*, December 2024, 1-22.

Metz Cade (2021). *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World*. New York: Dutton.

Naik Dishita, Naik Ishita, Naik Nitin (2024). "Large Data Begets Large Data: Studying Large Language Models (LLMs) and Its History, Types, Working, Benefits and Limitations". In: *Contributions Presented at The International Conference on Computing, Communication, Cybersecurity and AI, July 3–4, 2024, London, C3AI 2024.* Cham: Springer, 239-314. https://doi.org/10.1007/978-3-031-74443-3_18

Raus Rachele (2024). "*Deep learning* e traduzione intralinguistica: riformulare i testi della pubblica amministrazione in modo inclusivo". *Studi italiani di Linguistica Teorica e Applicata*, LII, 3, 350-367.

Raus Rachele, Tonti Michela, Cerquitelli Tania, Cagliero Luca, Attanasio Giuseppe, La Quatra Moreno, Greco Salvatore (2022). "L'analyse du discours et l'intelligence artificielle pour réaliser une écriture inclusive : le projet E-MIMIC". In: *8e Congrès Mondial de Linguistique Française,* HS Web Conf, vol. 138, 1-15. https://www.shs-conferences.org/articles/shsconf/pdf/2022/08/shsconf_cmlf2022_01007.pdf

Sabatini Alma (1987). *Raccomandazioni per un Uso Non Sessista Della Lingua Italiana*. Commissione nazionale per la realizzazione della parità tra uomo e donna, Presidenza del Consiglio dei Ministri.

Saussure Ferdinand de (1916), *Cours de linguistique générale*. Paris: Payot.

Yvon François (2023). "La traduction multilingue: analyse d'une prouesse technologique". *MediAzioni*, 39, A17–A34.

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

Veale Tony (2021). *Your Wit Is My Command: Building AIs with a Sense of Humor*. Cambridge: The MIT Press.